# UNIVERSITY OF SOUTHAMPTON

FACULTY OF BUSINESS, LAW AND ART

Web Science

**Academic Research Data Re-usage in a Digital Age: Modelling Best Practice**

by

**Laura Elizabeth German**

Thesis for the degree of Doctor of Philosophy

October 2015

**UNIVERSITY OF SOUTHAMPTON**

# ABSTRACT

FACULTY OF BUSINESS, LAW AND ART

Web Science

Thesis for the degree of Doctor of Philosophy

## ACADEMIC RESEARCH DATA RE-USAGE IN A DIGITAL AGE: MODELLING BEST PRACTICE

by Laura Elizabeth German

Recent high profile retractions – such as the case of Woo Suk Hwang and others – demonstrate that there are still significant issues regarding the reliability of published academic research data. While technological advances offer the potential for greater data re-usability on the Web, models of best practice are yet to be fully re-purposed for a digital age.

Employing interdisciplinary web science practices, this thesis asks what makes for excellent quality academic research across the sciences, social sciences and humanities. This thesis uses a case study approach to explore five existing digital data platforms within chemistry, marine environmental sciences and modern languages research. It evaluates their provenance metadata, legal, technological and socio-cultural frameworks. This thesis further draws on data collected from semi-structured interviews conducted with eighteen individuals connected to these five data platforms. The participants have a wide range of expertise in the following areas: data management, data policy, academia, law and technology.

Through the interdisciplinary literature review and cross-comparison of the three case studies, this thesis identifies the five main principles for improved modelling of best practice for academic research data re-usage both now and in the future. These principles are: (1) sustainability, (2) working towards a common understanding, (3) accreditation, (4) discoverability, and (5) a good user experience. It also reveals nine key grey areas that require further investigation.

# Table of Contents

# List of tables



# List of figures

# DECLARATION OF AUTHORSHIP

I, Laura Elizabeth German

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Academic Research Data Re-usage in a Digital Age: Modelling Best Practice

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

   Presented an extended abstract entitled: 'Overcoming Methodological Silos through Interdisciplinary Research: The Case of the Web Scientist' at the London Centre for Social Studies PhD Conference 2013: Methodological Choices and Challenges. King's College London. 19 April 2013. **Unpublished.**

   Presented an abstract entitled: 'How the Law Supports Existing Models of Environmental Data Reuse: The 'Marine Environmental Data and Information Network' (MEDIN) Case Study' at 'Information and Communications Technology for Environmental Regulation: Developing a Research Agenda Workshop'. The National University of Ireland, Galway. 20 – 21 June 2013. **Unpublished.**

   Laura German, 'MEDIN Case Study: PhD Research Accepted by the 'ICT for Environmental Regulation Workshop'', Marine Data News, 24 June 2013 <http://medin.newsweaver.co.uk/marinedata/16had7wvxub> [accessed 2 November 2014].

Signed: Laura Elizabeth German

Date: 16 October 2015

# Acknowledgements

I would like to express my sincere gratitude to my lead supervisor Professor Mary Orr. Thank you for all your enthusiasm, support, extensive feedback and our many wonderful discussions during the PhD. I have learnt so much from you, and have been grateful for the opportunity to continue working together during the post-doctoral project.

Thank you to my co-supervisor Professor Stephen Saxby for furthering my knowledge of IT law. Further thanks to my other co-supervisor Professor Leslie Carr for introducing me to the fascinating area of open access and open data.

I would also like to thank the eighteen individuals who so generously gave their time to participate in semi-structured interviews. Thanks to Helen Campbell who helped me arrange the meeting in Liverpool.

I would not have embarked on this Web Science PhD without Caroline Wilson to whom I owe my passion for intellectual property law – an inspirational lady who encouraged me to apply to the first Web Science Cohort at the University of Southampton.

A special thanks goes to the Web Science DTC directors for the amazing research opportunities during the PhD, and to Claire Wyatt and Dani O'Shea for all their fantastic assistance. I gratefully appreciate the funding from EPSRC as part of the Digital Economy Research Councils UK initiative.

My deepest gratitude extends to Professor John Coggon for all his help during the final stages leading up to submission, and Professor Nicholas Hopkins for his support during the upgrade. Many thanks also to Debbie Evans at the PGR office for answering all my queries, Dr Sophie Stalla-Bourdillon for her valuable comments during the upgrade, and Roksana Moore for our useful discussions during the initial stages of the thesis.

Acknowledgement goes to my two proof-readers – Adam Carmichael and Christopher German. Further thanks to my mum, Elizabeth and my dad, Andrew for your comments.

Many thanks to my *viva voce* examiners – Professor Hector MacQueen and Dr Nicholas Gibbins – for your insightful comments and analysis.

This thesis would not have been possible without the understanding and encouragement from all my family and friends. A huge thank you goes to my mum, dad and brother for their unwavering support not only during the PhD but throughout my entire education. To my husband Adam – thank you for your constant encouragement and friendship. It has all meant so much.

I therefore dedicate this thesis with love to my mum, dad, brother and husband as an expression of my greatest appreciation. I could not have undertaken this process without you all.

# Acronyms

ACM = Association for Computing Machinery
ADS = Archaeology Data Service
AHRC = Arts and Humanities Research Council
AICC = Aviation Industry CBT Committee
ASCII = American Standard Code for Information Interchange
BAAL = British Association of Applied Linguistics
BBC = British Broadcasting Corporation
BBSRC = Biotechnology and Biological Sciences Research Council
BODC = British Oceanographic Data Centre
CC = Creative Commons
CD = Compact Disc
CDPA = Copyright, Designs and Patents Act 1988
CEFAS = Centre for Environment, Fisheries & Aquaculture Science
CERN = European Organisation for Nuclear Research
CHAT = Codes for the Human Analysis of Transcripts
CHILDES = Child Language Data Exchange System
CIF = Crystallographic Information File
CLAN = Computerised Language Analysis
CML = Chemical Mark-up Language
COPE = Committee on Publication Ethics
CRB = Criminal Record Bureau
CSV = Comma Separated Values
DAC = Data Archive Centre
DASSH = Data Archive for Marine Species and Habitats
DbCL = Database Contents License
DBS = Disclosure and Barring Service
DCAT = Data Catalog Vocabulary
DCC = Digital Curation Centre
DEFRA =Department for Environment, Food and Rural Affairs
DOI = Digital Object Identifier
ELN = Electronic Laboratory Notebook
EMODnet = European Marine Observation and Data Network
EPSRC = Engineering and Physical Research Council
EPSRC = Engineering and Physical Sciences Research Council
ES = Embryonic Stem
ESDS = Economic and Social Data Service
ESRC = Economic and Social Research Council
EU= European Union
FLLOC = French Learner Language Oral Corpora
HO = Hydrographic Office
HTML = Hypertext Mark-up Language
HTTP =Hypertext Transfer Protocol
INSPIRE Directive = Infrastructure for Spatial Information in Europe Directive
ISO = International Organization for Standardization
IT = Information Technology
JSON = JavaScript Object Notation
LANGSNAP = Languages and Social Networks Abroad Project
MEDIN = Marine Environmental Data and Information Network

MHRA = Modern Humanities Research Association
MRC = Medical Research Council
NCS = National Crystallography Service
NERC= Natural Environment Research Council
NOAA = National Oceanic and Atmospheric Administration
NOCS = National Oceanography Centre, Southampton
ODC = Open Data Commons
ODI = Open Data Institute
OECD = Organisation for Economic Co-Operation and Development
OED = Oxford English Dictionary
OGL = Open Government Licence
OKF = Open Knowledge Foundation
PDF = Portable Document Format
RAE = Research Assessment Exercise
RCUK = Research Councils UK
RDF = Resource Description Framework
REF = Research Excellence Framework
RIS = Research and Innovation Services, University of Southampton
SCCG = Southampton Chemical Crystallography Group
SPLLOC = Spanish Learner Language Oral Corpora
STFC = Science and Technology Facilities Council
UCML= University Council of Modern Languages
UK = United Kingdom of Great Britain and Northern Ireland
URI = Universal Resource Indicator
URL = Universal Resource Locator
USA = United States of America
USPTO = United States Patent and Trademark Office
W3C = World Wide Web Consortium
WIPO = World Intellectual Property Organization
XML = Extensible Mark-up Language

# Chapter 1: Introduction

## 1.1 Motivation for research

Across all disciplines in the sciences, social sciences and humanities, academic research activity and outputs in the UK university sector operate according to among the most robust self-monitoring systems in the world. The UK higher education sector is subject to a number of stringent protocols including: institutional and research council best practice guidance and policies; long-established peer review through conferences, academic publishers, and funding councils; and, scrutiny via institutional and learned society ethics boards. Together they qualify and establish excellence in UK academic research. Moreover, the UK is home to many of the oldest and most respected academic journals in the world; for example, *Philosophical Transactions* founded in 1665 and *Nature* founded in 1869 and renowned today.[1]

Despite all these stringent protocols and safeguards, instances of erroneous research and academic misconduct – from minor to major incidents – occur regularly. Bad practice may manifest through genuine mistakes and/or ignorance. For example, such practice could result in the (unintentional) loss, duplication, and/or open release of low quality academic research data. Worst practice occurs at the highest echelons of academic misconduct where research data are (often intentionally) manipulated, fabricated, unlawful, and/or unethical.

The case of Woo Suk Hwang and others is one such instance of worst practice, which led to the retraction of two stem cell research articles published in the authoritative journal *Science*. This case is used as a motivating example of worst practice in a digital age, and is critically examined in Chapter 2 of this thesis (see section 2.2). Therefore, while it may be clear what constitutes excellent academic research, it appears that this requires further re-defining and strengthening in a digital age.

[1] For a digitised version of the first issue of *Philosophical Transactions* refer to: *Philosophical Transactions*, 1 (1 January 1665), (1-22) <http://rstl.royalsocietypublishing.org/content/1/1-22.toc> [accessed 9 August 2015]; for a digitised version of the first issue of *Nature* refer to: *Nature*, 4 November 1869 <http://www.nature.com/nature/about/first/> [accessed 9 August 2015].

The technological architecture that comprised the foundations for the World Wide Web was first proposed by Tim Berners-Lee in March 1989:

> Tim Berners-Lee submitted a proposal for an information management system to his boss, Mike Sendall. *'Vague, but exciting'*, were the words that Sendall wrote on the proposal, allowing Berners-Lee to continue.'[2] (CERN's italicised emphasis.)

Berners-Lee worked at the European Organisation for Nuclear Research (CERN) where physicists from assorted organisations often faced difficulties sharing their data due to the multitude of different formats available.[3] The Web overcame these problems and offered a universally accepted method to share data:

> On 30 April 1993 CERN put the World Wide Web software in the public domain. CERN made the next release available with an open licence, as a more sure way to maximise its dissemination. Through these actions, making the software required to run a web server freely available, along with a basic browser and a library of code, the web was allowed to flourish.[4]

A knowledge explosion ensued, as individuals and organisations established websites to share their data. Today, the Web is used by various sectors of society from academia and the media to e-commerce and social networks. Twenty years after the worldwide release of the Web, this thesis focuses on its original aim – to facilitate data sharing within the academic research community – according to existing best practices for academic research data re-usage to discover how such practices can be better developed in the digital age.

The Web has enabled data sharing to flourish through its instantaneous speed of transfer, its vast and increasing storage capacity, its potential global reach, ease of search (engines), and facility for copy, cut and paste. In a digital age, academic research data are no longer confined by print and restrained by physical delivery; data have the potential to be immediate, mutable and reach wider audiences than ever before. While

---

[2] Quoted from: 'Tim Berners-Lee's Proposal', *European Organisation for Nuclear Research (CERN) Website* <http://info.cern.ch/Proposal.html> [accessed 9 August 2015]; for further information refer to: Tim Berners-Lee, 'Information Management: A Proposal', *Original Proposal for a Global Hypertext Project at CERN, W3 Website* (1989) <http://www.w3.org/History/1989/proposal.html> [accessed 9 August 2015]; 'Tim Berners-Lee', *W3 Website* <http://www.w3.org/People/Berners-Lee/> [accessed 9 August 2015].

[3] 'The birth of the Web', *European Organisation for Nuclear Research' (CERN) Website* <http://home.web.cern.ch/topics/birth-web> [accessed 9 August 2015].

[4] *Ibid.*

good quality data are valued as the building blocks of academic analysis and interpretation, the digital age marks a 'change in value' of these data.[5]

Due to the increased computational nature of academic research, the growth in the amount of data produced has been unprecedented.[6] In a digital age, data are viewed as an integral part of 'the lifeblood of a knowledge-based economy.'[7] For Renée Marlin-Bennett 'the "new" economy depends on buying and selling ideas and facts, intangible and ephemeral though they are'.[8] A significant proportion of academic research data are without direct commercial value however. Their value lies in contributing to future research, policy decisions and above all furthering human knowledge and understanding.

Across all disciplines in the sciences, the social sciences and the humanities, the academic community are using Web technologies on an unprecedented scale to make their data openly accessible to a global audience. In the digital age, academic research data have emerged as their own entity, as Catherine Colston summarises:

> The digital revolution has rendered information and its collation an asset of increased significance and value. Collection, arrangement and presentation of information is indispensable to business and financial services, government, the scientific and educational communities, and to consumers. It is a new industry in itself. [9]

Long-held principles of research integrity are becoming ever more important, and challenging, to uphold within a myriad of (often assimilated) data. Academic research data still require validation, quality assurance, preservation, ethics approval, rights clearance, and rights management.

In consequence, a key motivation for this thesis is to examine the re-usage of academic research data: for (1) how it has been re-versioned for a digital age; and, (2)

---

[5] Stephen Saxby, *The Age of Information* (London: The Macmillan Press Ltd, 1990), p. 1.

[6] Martin Hilbert and Priscila López, 'The World's Technological Capacity to Store, Communicate, and Compute Information', *Science*, 322 (6025) (2011), 60-65 <http://dx.doi.org/10.1126/science.1200970>; Jon Stewart, 'Global data storage calculated at 295 exabytes', *BBC News Online*, 11 February 2011 <http://www.bbc.co.uk/news/technology-12419672> [accessed 9 August 2015]; Yochai Benkler, *The Wealth of Networks: How Social Production Transforms Markets and Freedoms* (London: Yale University Press, 2006).

[7] Peter K. Yu, (ed.) *Intellectual Property and Information Wealth: Issues and Practices in the Digital Age* (Connecticut, USA: Greenwood Publishing Group, 2007), p. ix. Google eBook.

[8] Renée Marlin-Bennett, *Knowledge Power: Intellectual Property, Information and Privacy* (London, Boulder, 2004), p. 1.

[9] Catherine Colston, '*Sui Generis* Database Right: ripe for review?' *Journal of Information, Law and Technology*, (3) (2001) <http://strathprints.strath.ac.uk/629/> [accessed 9 August 2015]. ePrint.

what best practice principles are required to guarantee its robustness across the sciences, social sciences and humanities both now and in the future.

## 1.2 Academic research data definition

There are two types of data produced across the sciences, the social sciences and the humanities: quantitative and qualitative. The Oxford English Dictionary (OED) defines data as:

> Facts and statistics collected together for reference or analysis; the quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media; *philosophy* things known or assumed as facts, making the basis of reasoning or calculation.[10]
> (OED's italicised emphasis.)

The dictionary definition above focuses on quantitative data and overlooks qualitative data. Quantitative data are numerical; for example statistics, numerical measurements and graphs, usually associated with scientific endeavour. Keith F. Punch states that quantitative data commonly 'are in the form of numbers' whereas qualitative data 'are not in the form of numbers'.[11] Qualitative data are non-numerical for example case law, statutes, memoirs, bibliographies, interview sound recordings and verbatim transcriptions. This thesis defines data as follows (refer to glossary):

> All qualitative (non-numerical) and quantitative (numerical) materials, which include: facts, observations, measurements, statistics, figures, lists, sound recordings, verbatim transcripts, bibliographies, memoirs, case law and statutes; materials obtained for interpretation and analysis to produce information, knowledge and/or wisdom.

It further defines academic research data (refer to glossary):

> Any qualitative or quantitative data that have the potential for (re)use, either partially or completely that are produced in the course of academic research within the sciences, social sciences and humanities.

Data are the fundamental building blocks upon which knowledge claims are then based, and made into academic journal articles, monographs, reports, books and other scholarly works. The following example, taken from Gene Bellinger and others, illustrates that it is only through the aggregation of raw data that further data, information, knowledge,

---

[10] 'Definition: Data', *Oxford Dictionaries Website* <http://oxforddictionaries.com/definition/data> [accessed 9 August 2015].

[11] Keith F. Punch, *Introduction to Social Research: Quantitative and Qualitative Approaches*, 2nd edn (London, SAGE Publications 2005), p. 3. Google eBook.

and, in some cases, wisdom are created.[12] If each bullet point represents a raw datum, it has little meaning on its own:

> I have a box.
>
> • The box is 3' wide, 3' deep, and 6' high.
>
> • The box is very heavy.
>
> • The box has a door on the front of it.
>
> • When I open the box it has food in it.
>
> • It is colder inside the box than it is outside.
>
> • You usually find the box in the kitchen.
>
> • There is a smaller compartment inside the box with ice in it.
>
> • When you open the door the light comes on.
>
> • When you move this box you usually find lots of dirt underneath it.
>
> • Junk has a real habit of collecting on top of this box.
>
> What is it? [/] A refrigerator. You knew that, right? At some point in the sequence you connected with the pattern and understood it was a description of a refrigerator. From that point on each statement only added confirmation to your understanding. [/] If you lived in a society that had never seen a refrigerator you might still be scratching your head as to what the sequence of statements referred to.[13] (Bellinger and others' bullet points.)

The analysis and interpretation of erroneous or incomplete academic research data ultimately leads to poor quality articles, books, monographs, reports and other scholarly works. For instance if the only data provided above were 'the box is very heavy' and 'the box has a door on the front of it', it would be more difficult to arrive at the conclusion that the box is a refrigerator. The validation of academic research data as robust, complete and without flaws is obviously essential to its re-usability in many foreseen and unforeseen forms. In consequence, the prevention and detection of erroneous academic research data are of vital importance across the sciences, social sciences and humanities.

---

[12] Gene Bellinger, Durval Castro and Anthony Mills, 'Data, Information, Knowledge, and Wisdom' *Systems Thinking Online Article* (2004) *Mental Model Musings Website* <http://www.systems-thinking.org/dikw/dikw.htm> [accessed 9 August 2015].

[13] Bellinger and others.

## 1.3 Research questions and thesis contributions

Many disciplines have raised a wide-range of issues concerning the robustness, reliability and re-usability of academic research data. Despite this extensive focus, there is currently no foremost authority that wholly encapsulates: (a) how the re-usage of academic research data has been re-versioned for a digital age; and, (b) what best practice principles are required to guarantee its robustness across the sciences, social sciences and humanities both now and in the future. This thesis is therefore located in the interdisciplinary field of web science to bring together the crucial knowledge bases and methodologies required to address the following primary research questions:

1) What makes for excellent quality academic research in a digital age?
2) How should best practice for academic research data re-usage be modelled both now and in the future? What can be learnt from longstanding practices?

In consequence, the literature review in Chapter 2 demonstrates that the key literature is often disjointed and scattered within a number of disciplinary domains. By joining up and filling the gaps between these previously unconnected areas in ways which have not yet been achieved, this thesis offers a new synthesis of existing research. The methodological and interdisciplinary approach employed by this thesis is therefore a significant contribution in itself.

Given the interdisciplinary field of web science, this thesis draws on three disciplinary domains: the humanities, law, and web and internet science. A humanities approach is used to unpack the longstanding principles that continue to scope the re-usage of academic research data in a digital age, such as peer-review, data accessibility, trust, authorisation and traceability. A legal approach focuses on some of the key UK legal issues impacting on the authorised collection, management and re-usage of academic research data in practice with significant emphasis on licensing. A web and internet science approach examines crucial IT issues affecting academic research data re-usage in a digital age. This thesis is therefore written for an interdisciplinary audience and co-supervised by an academic from each of these three areas. The literature review utilises an extensive range of materials from a multitude of disciplines.

Moreover, the case of Hwang and others, also in Chapter 2, is selected as a motivating example of worst practice: (a) to emphasise the underlying interdisciplinary frameworks that are required to support the generation, management and re-usage of academic research data in a digital age; (b) to highlight particular areas for further attention where worst practice research was able to pass through the longstanding peer-

review and other assurance processes that should have operated to halt its further dissemination and re-usage; and, (c) to better understand best practice, as without first understanding worst practice it is difficult to measure its scope for use in preventing and immobilising minor incidents of bad practice as well as those major incidents of worst practice.

In addition to the two primary research questions, this thesis identifies three secondary research questions concerning legal, socio-cultural, ethical and technological dimensions of research through the motivating example of worst practice:

3) How should diverse types of academic research data from multiple originators, contexts and sources be safeguarded for a wide set of research users?
4) How should academic research data generated as part of collaborative research be treated in order to balance a range of difficult communal and continuity problems?
5) How should maintenance of and access to sensitive and personalised data be treated in order to balance a range of difficult problems with permissions, data protection and confidentiality?

While the development of these research questions are explained in detail within Chapter 2, these secondary research questions are used to sharpen focus on the main contributions of this thesis and to further facilitate the discovery of key grey areas for future study.

This thesis recognises that there is not only an intellectual need but pragmatic requirement, to critically assess how to most effectively model best practice for academic research data re-usage in a digital age. A case study approach is therefore employed, in Chapters 4-6, to explore five existing digital data platforms across usually separated disciplines within different university faculties: chemistry, marine environmental sciences and modern languages. These three case studies deliberately lie outside the thesis author's prior higher education experience to prevent pre-given assumptions concerning academic research data re-usage.

This case study approach is used to evaluate how these data platforms determine and promote quality data generation, management and re-usage at all stages of the research process. This thesis evaluates their provenance, legal, technological and socio-cultural frameworks. It further draws on data collected from semi-structured interviews conducted with eighteen individuals connected to these five data platforms. The participants have a wide range of expertise in the following areas: data management, data policy, academia, law and technology. This methodological approach is fully outlined within Chapter 3.

In direct response to these research questions, the two most important contributions this thesis offers are: (a) the identification of five best practice principles (through three case study chapters) for effective and improved academic research data re-usage across the sciences, social sciences and humanities both now and in the future (discussed in Chapter 7); and, (b) nine key grey areas for future work (set out in Chapter 8).

## 1.4 Thesis summary

Chapter 1 has introduced the key motivation for this thesis and the research questions to be addressed. Chapter 2 will provide an interdisciplinary literature review and begin to identify significant narrative threads pertinent to the questions to be investigated in the case studies. Chapter 3 outlines and justifies the interdisciplinary methodological approach utilised by the thesis. Chapters 4-6 focus on three distinct case studies within the marine environmental sciences, chemistry and modern languages. These case studies include critical analysis of primary source materials and semi-structured interviews. Chapter 7 details the five best practice principles for effective academic research data re-usage across the sciences, social sciences and humanities both now and in the future. Finally, Chapter 8 completes the thesis by outlining its conclusions and significant areas for future work.

# Chapter 2: Literature review

## 2.1 Literature review: Introduction

This chapter opens by examining a high profile case of worst practice centred on two retracted articles published within the authoritative journal *Science* by Woo Suk Hwang and others in 2004-5 (see section 2.2). This is used as a motivating example to map how the safeguards which should have prevented the publication of these two articles failed. This serves as a preamble to the literature review (see section 2.3) which examines the contextual and historical development of these safeguards from non-digital (print era) to digital (e-print era) research dissemination. It examines the four core elements of these safeguards – (1) data accessibility, (2) trust, (3) authorisation, and (4) traceability – which add up to good quality academic research data re-usage. This section further introduces a number of significant narrative threads located throughout Chapters 4-6 and further maps these interdisciplinary issues to these three case study chapters.

## 2.2 Bad and worst practice in a digital age

Bad practice may manifest through genuine mistake and/or ignorance. The three case studies (see Chapters 4-6) identify a number of commonplace bad practice examples, which occur across academic research domains. These examples include where academic research data are lost, because there is no secure archival available (Chapter 4) and/or projects are shelved without maintaining the data (Chapter 5). In some cases where data are collected from human participants insufficient consent is obtained; therefore, preventing the future release of the data (Chapter 6). Academic research data may be (openly) released in non-standard formats without (or with insufficient) information about their authorisation, attribution, history, quality, peer review, versioning, formats, context and/or other vital aspects of their lineage.

The case of Woo Suk Hwang transcends these more common incidents of bad practice and therefore offers an extreme example of worst practice. The five principles for improved modelling of best practice recommended by this thesis (see Chapter 7) not only aim to cover the minor incidents of worst practice, but to further help prevent and immobilise major incidents. In consequence, the case of Hwang and others is used as a

motivating example to probe some of the more serious issues with the safeguards that are meant to ensure good quality academic research data re-usage.

### 2.2.1 Academic journal publication

Publication within academic journals still remains one of the hallmarks of academic excellence. However, the main focus for scrutiny has long been the academic analysis and interpretation rather than the underlying and supporting academic research data leading to them. This situation is gradually changing with the increase in data journals. Sarah Callaghan provides a non-exhaustive list of fifteen existing data journals or journals which publish data papers (this list was last updated on 22 January 2013): *Geoscience Data Journal, Earth System Science Data, Ecological Archives, Dataset Papers in Science, Journal of Chemical and Engineering Data, GigaScience Journal, Journal of Physical and Chemical Reference Data, F1000 Research, International Journal of Robotics, CODATA Data Science Journal, Journal of Open Archaeology Data, Journal of Open Public Health Data, Journal of Open Psychology Data, Journal of Open Psychology Data,* and *BMC Research Notes.*[14] However, this list does not

[14] Refer to: Sarah Callaghan, 'A list of Data Journals (in no particular order)', *trac Website* <http://proj.badc.rl.ac.uk/preparde/blog/DataJournalsList> [accessed 9 August 2015]; *Geoscience Data Journal* (first issue published in June 2014), *Wiley Website* <http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060> [accessed 9 August 2015]; *Earth System Science Data* (first issue 2009), *Earth System Science Data Website* <http://www.earth-system-science-data.net/> [accessed 9 August 2015]; Data papers can be submitted to *Ecological Archives*, *Ecological Society of America Website* <http://esapubs.org/archive/instruct_d.htm> [accessed 9 August 2015]; *Dataset Papers in Science*, *Hindawi Publishing Corporation Website* <http://www.hindawi.com/journals/dpis/> [accessed 9 August 2015]; *Journal of Chemical and Engineering Data*, *ACS Publications Website* <http://pubs.acs.org/journal/jceaax> [accessed 9 August 2015]; *GigaScience Journal*, *GigaScience Journal Website* <http://www.gigasciencejournal.com/> [accessed 9 August 2015]; *Journal of Physical and Chemical Reference Data* (first issue published in 1972), *AIP Scitation Website* <http://scitation.aip.org/content/aip/journal/jpcrd/browse> [accessed 9 August 2015]; *Biodiversity Data Journa*l, *Biodiversity Data Journal Website* <http://biodiversitydatajournal.com/> [accessed 9 August 2015]; *F1000 Research*, *F1000 Research Website* <http://f1000research.com/> [accessed 9 August 2015]; *International Journal of Robotics* (publishes data papers), *SAGE Journals Website* <http://ijr.sagepub.com/> [accessed 9 August 2015]; *CODATA Data Science Journal*, *CODATA Website* <http://www.codata.org/publications/data-science-journal> [accessed 9 August 2015]; *Journal of Open Archaeology Data*, *Journal of Open Archaeology Data Website* <http://openarchaeologydata.metajnl.com/> [accessed 9 August 2015]; *Journal of Open Public Health Data*, *Journal of Open Public Health Data Website* <http://openhealthdata.metajnl.com/>

include the PLOS ONE journals, which data policy states: 'PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception'.[15] Furthermore, since this list was last updated, *Nature* also launched a new data journal in 2014 called *Scientific Data*:[16]

> Data Descriptors will link to both related journal articles and data files stored at data repositories, helping readers to navigate easily between research, data descriptions and the actual data. And each Data Descriptor publication will be supported by machine-readable experimental metadata to help advanced users mine and search *Scientific Data*'s content. […] [|][…][|] What *Scientific Data* will not be is a new data repository. Rather, it will promote and cooperate with existing community-based repositories, and will combat data fragmentation by ensuring that data sets are deposited in an appropriate repository.[17]

As data journals only form a minority of the academic publication sector, a considerable number of academic research data are not directly and/or thoroughly reviewed as part of the publication process. This lack of scrutiny is problematic and has contributed to a significant number of retractions, particularly in scientific publications.

In an age where mass distribution of academic research data is now possible, many academic publishers have only sought to re-version articles for a digital media. While the majority of academic articles are now digitally accessible, such articles are published within the parameters of print media. Many journals therefore have not sought to capture and make the underlying and supporting academic research data accessible.

As a high profile example of academic misconduct, the case of Woo Suk Hwang and others is well-placed to address these worst practice issues. The following section focuses on how the procedural issues that led to academic articles built on erroneous data being published by the authoritative journal, *Science*, were overcome. Moreover, it reveals how such academic misconduct proved sufficiently serious to necessitate legal action, and cause detriment to a national economy.

---

[accessed 9 August 2015]; *Journal of Open Psychology Data*, *Journal of Open Psychology Data Website* <http://openpsychologydata.metajnl.com/> [accessed 9 August 2015]; *BMC Research Notes* (publishes datasets), *BioMed Central Website* <http://www.biomedcentral.com/bmcresnotes/> [accessed 9 August 2015].

[15] 'PLOS Editorial and Publishing Policies', *PLOS ONE Website* < https://www.plos.org/policies/> [accessed 9 August 2015].

[16] *Scientific Data, Nature Website* <http://www.nature.com/sdata/> [accessed 9 August 2015].

[17] Announcement: Launch of an online data journal', *Nature*, 502, (142) (10 October 2013) <http://dx.doi.org/10.1038/502142a>

While it must be made clear that there are a number of significant cases of academic misconduct that occur within the UK and other countries around the world, the case of Woo Suk Hwang and others was selected as a well-known illustration of worst practice with a multitude of thought-provoking issues around data accessibility, trust, authorisation, traceability, and its many breaches of national and institutional protocols for quality academic research.

### 2.2.2 The case of Woo Suk Hwang and others

In 2004-5, two articles written by Hwang and others in the authoritative journal *Science* made claims that significantly advanced stem cell research.[18] Their scientific claims are described by Gretchen Vogel:

> Woo Suk Hwang and his colleagues told the world that they could make embryonic stem (ES) cells from cloned human embryos with an efficiency that astounded – and thrilled – their colleagues. In roughly one out of every 12 tries, the South Korean team reported, they could produce ES cell lines that were a genetic match to patients. Scientists hoped to use such cells to probe the genetic triggers of diseases such as diabetes and amyotrophic lateral sclerosis (ALS). Some dreamed of using them as raw material for developing new tissues and cells that could treat previously incurable maladies.[19]

Woo Suk Hwang – at this time a Professor at Seoul National University – led the innovative research team and received 'rock-star status' in South Korea for his work on cloning.[20] The research team also received a mention in the 2004 *TIME* 100 list of people who matter.[21] These scientific claims were found to be fraudulent however. The embryonic stem cell lines had never been produced. The academic research data were found to be fabricated, falsified and unethical.

---

[18] Woo Suk Hwang and others, 'Evidence of a Pluripotent Human Embryonic Stem Cell Line Derived from a Cloned Blastocyst', *Science*, 303 (5664) (2004), 1669-1674 <http://dx.doi.org/10.1126/science.1094515> ; Woo Suk Hwang and others, 'Patient-Specific Embryonic Stem Cells Derived from Human SCNT Blastocysts', *Science*, 308 (5729) (2005), 1777-1783 <http://dx.doi.org/10.1126/science.1112286>

[19] Gretchen Vogel (with reporting by Dennis Normile), 'NewsFocus: Picking Up the Pieces After Hwang', *Science*, 312 (5773) (28 April 2006), 516-7 (p.516) <http://dx.doi.org/10.1126/science.312.5773.516>

[20] Constance Holden (with reporting by Gretchen Vogel and Dennis Normile), 'News: Korean Cloner Admits Lying About Oocyte Donations', *Science*, 310 (5753) (2 December 2005), 1402-1403 <http://dx.doi.org/10.1126/science.310.5753.1402>

[21] Jeffrey Kluger, 'Scientists and thinkers: the Korean cloners', *TIME*, 26 April 2004 <http://www.time.com/time/specials/packages/article/0,28804,1970858_1970909_1971678,00.html> [accessed 9 August 2015].

In May 2004, the authoritative journal *Nature* questioned whether the academic research data were ethically produced, because there were allegations that donors were paid for their eggs and some of Hwang's researchers were donors.[22] This did not amount to an investigation however. On 1 June 2005, an individual, claiming to have worked with Hwang on the research that was published in the 2004, sent an anonymous message to an investigative South Korean television programme called PD Notebook.[23] The whistle-blower claimed to have left Hwang's research team before publication of the 2004 article, due to concerns over ethical violations and technical issues. Although the individual did not have any physical evidence to support the claims, PD Notebook began to investigate. However, the investigation was hampered by their unethical journalistic practices. Whilst interviewing Hwang's co-author on camera – Sun Jong Kim – they pretended to have physical evidence of the fraudulent claims in their possession. They therefore deceived Kim to believe that a police investigation had begun, and did not respond when asked whether they had stopped filming.[24]

In December 2005, anonymous posts on research message boards – such as the Biological Research Information Center – by the young South Korean academic community then cast significant doubts over the validity of the academic research data, including the DNA fingerprints and images.[25] On the 12 January 2006 *Science* released a formal retraction of these two articles.[26] In 2004-5, Hwang and his colleagues had also published articles in nine other journals. Many including *Molecular Reproduction and*

---

[22] David Cyranoski (with additional reporting by Erica Check), 'Who's who: a quick guide to the people behind the Woo Suk Hwang story', *Nature News*, 11 January 2006 <http://dx.doi.org/10.1038/news060109-9>

[23] Sei Chong and Dennis Normile (with reporting by Gretchen Vogel), 'News of the Week: How Young Korean Researchers Helped Unearth a Scandal ...' *Science*, 311 (5757) (6 January 2006), 22-25 <http://dx.doi.org/10.1126/science.311.5757.22>

[24] *Ibid.*

[25] David Cyranoski (with additional reporting by Erica Check), 'Who's who: a quick guide to the people behind the Woo Suk Hwang story', *Nature News*, 11 January 2006 <http://dx.doi.org/10.1038/news060109-9>; *The Biological Research Information Center Website* <http://bric.postech.ac.kr/myboard/read.php?id=267&Board=bric_board> [accessed 9 August 2015].

[26] Donald Kennedy, 'Editorial Retraction: Retraction of Hwang *et al.*, *Science* 308 (5729) 1777-1783', *Science*, 311 (5759) (12 January 2006), 335 <http://dx.doi.org/10.1126/science.1124926>; Erika Check, 'Journals scolded for slack disclosure rules', *Nature News*, 18 January 2006 <http://dx.doi.org/10.1038/news060116-6>

*Development*, *Stem Cells* and *Theriogenology* also investigated potential misconduct.[27] This scandal not only caused 'profound shock' to the scientific community and a lapse in public support for embryonic research, but had a direct impact on the stock prices of South Korea's biotech industry.[28]

On 13 December 2005, prior to the retraction of Hwang and others' two articles in *Science*, Gerald Schatten, co-author with Hwang and others of the 2005 article and American biologist at the University of Pittsburgh, sought to remove his name from the paper. He argued that, on further analysis of the published paper and its data, their reliability could not be verified. *Science* had no procedure in place for authorial retraction of one author from a multi-authored submission.[29] The University of Pittsburgh conducted a panel to investigate Schatten's potential academic misconduct. The panel held that this resulted from a lack of responsibility to verify the academic research data and that Schatten had not taken part in the practical experimentation.[30] Therefore, Schatten was not in the position to be named as a senior co-author, as he had not had a considerable role in the research.

In January 2006, Woo Suk Hwang and his colleagues were accused of fraud, embezzlement, destruction of evidence and violations of bioethics law.[31] On 26 October 2009, the Seoul Central District Court held that Hwang was guilty of embezzling 830 million *won* of government funds, and of purchasing human eggs in breach of bioethics law, but found him not guilty of fraud.[32] He received a two-year prison sentence, suspended for three years.[33] Throughout the trial, Hwang maintained that he was

---

[27] Jennifer Couzin, Constance Holden and Sei Chong, 'News of the Week: Hwang Aftereffects Reverberate at Journals', *Science*, 311 (5759) (20 January 2006) 321 <http://dx.doi.org/10.1126/science.311.5759.321b>

[28] Ichiko Fuyuno, 'Business: Hwang scandal hits Korean biotech hard', *Nature*, 439 (19 January 2006), 265 <http://dx.doi.org/10.1038/439265a>; Mukta Jhalani, 'Protecting Egg Donors and Human Embryos – The Failure of the South Korean Bioethics and Biosafety Act', *Pacific Rim Law and Policy Journal*, 17 (3) (2008), 707-733 (p. 713).

[29] Erica Check, 'Stem-cell scientist asks for retraction: US partner urges Korean cloner to retract landmark paper', *Nature News*, 14 December 2005 <http://dx.doi.org/10.1038/news051212-5>

[30] Nicholas Wade, 'University Panel Faults Cloning Co-Author', *The New York Times*, 11 February 2006 <http://www.nytimes.com/2006/02/11/science/11clone.html> [accessed 9 August 2015].

[31] D. Yvette Wohn and Dennis Normile, 'Prosecutors Allege Elaborate Deception and Missing Funds', *Science*, 312 (5776) (19 May 2006), 980-981 <http://dx.doi.org/10.1126/science.312.5776.980>

[32] David Cyranoski, 'Woo Suk Hwang convicted, but not of fraud', *Nature News*, 461 (1181) (26 October 2009) <http://dx.doi.org/10.1038/4611181a>

[33] *Ibid.*

deceived by his junior researchers to believe that his team were triumphant in the production of embryonic stem cells from cloned human embryos.[34] However, it was confirmed that Hwang and his colleagues had successfully cloned a dog.[35]

Hwang was dismissed from his post at Seoul National University, yet on 19 August 2006 Hwang and his team continued their research into: 'animal cloning, animal stem cells, research on production of animal organs, and biological textile products'[36] publically at an independent laboratory in the new Suam Bioengineering Research Institute.[37] At the point of its creation, the Suam Bioengineering Research Institute made no reference to human cloning. In South Korea, human stem cell research must have government approval – Hwang's authorisation was withdrawn when the academic misconduct allegations were made.[38] In 2011 he alleged to have cloned the first coyotes and in 2012, Hwang and others were attempting to clone a woolly mammoth.[39]

Unfortunately, the academic misconduct of Hwang and his colleagues is not an isolated incident.[40] A former member of Hwang's research team – Jong Hyuk Park –

---

[34] Choe Sang-Hun, 'Researcher who faked cloning data gets new job – Asia – Pacific – International Herald Tribune', *The New York Times*, 18 August 2006 <http://www.nytimes.com/2006/08/18/world/asia/18iht-clone.2528877.html> [accessed 9 August 2015].

[35] Byeong Chun Lee and others, 'Dogs cloned from adult somatic cells', *Nature*, 436 (4 August 2005), 641 <http://dx.doi.org/ 10.1038/436641a>;
'Disgraced S Korean cloner Hwang back with Coyote Claim', *BBC News*, 17 October 2011 <http://www.bbc.co.uk/news/world-asia-pacific-15340240> [accessed 9 August 2015].

[36] Yudhijit Bhattacharjee (ed.,) 'Newsmakers: Movers – A Second Chance', *Science*, 313 (1 September 2006), 1233.

[37] Kim Rahn, 'Will Hwang Woo-Suk Return? Decision on Permits of Stem Cell Research Due Sat', *The Korean Times Online*, 27 July 2008 <http://www.koreatimes.co.kr/www/news/nation/2010/09/117_28274.html> [accessed 9 August 2015].

[38] Choe, Sang-Hun, 'Researcher who faked cloning data gets new job – Asia – Pacific – International Herald Tribune'.

[39] 'Disgraced S Korean cloner Hwang back with Coyote Claim', *BBC News*; 'Mammoth Task: Plan to Clone Ice Age Beast', *Sky News*, 13 March 2012 <http://news.sky.com/story/2931/mammoth-task-plan-to-clone-ice-age-beast> [accessed 9 August 2015]; 'South Korean and Russian scientists bid to clone mammoth', *Telegraph* , 13 March 2012 <http://www.telegraph.co.uk/earth/wildlife/9139976/South-Korean-and-Russian-scientists-bid-to-clone-mammoth.html> [accessed 9 August 2015].

[40] Phil Baty, 'Leader: Box ticked, but job not yet done', *The Times Higher Education*, 12 July 2012 <http://www.timeshighereducation.co.uk/story.asp?storycode=420544> [accessed 9 August 2015]; Debora Weber-Wulff, 'Viewpoint: The spectre of plagiarism haunting Europe', *BBC News*, 25 July 2012 <http://www.bbc.co.uk/news/18962349> [accessed 9 August 2015].

was later alleged to have fabricated monkey cloning data.[41] There are also many other examples of alleged academic misconduct, including: Jan Hendrik Schön – 'fabricated and falsified research findings'; Alexander Kugler – 'lack of proper data handling and record keeping'; and Elias Alsabri – 'fraud; plagiarism'.[42] Blogs – including: Retraction Watch, Embargo Watch, and Copy, Shake and Paste: A Blog about Plagiarism and Scientific Misconduct – are kept by individuals involved with academic journal publication to offer a record of past and current information about academic misconduct (allegations).[43]

In two recent articles, published in *PloS Medicine* by Daniele Fanelli and *PloS One* by R. Grant Steen and others, the number of retractions particularly in the sciences appears to be growing.[44] However, from these articles the reasons behind this growth are not entirely clear and not all retractions stem from academic misconduct: often genuine errors, mistakes and flaws may be the reason. For the Hwang case, the main failings of the safeguards can be reduced down into three main areas for concern: (1) the underlying and supporting data were not made available to the peer-reviewers and therefore not subsequently scrutinised – there were also allegations of data destruction that prevented any checks after the alleged misconduct; (2) the roles and contributions of the authors were not explicit or verified – i.e. Schatten did not have any part in the

---

[41] Constance Holden, 'News of the Week: Former Hwang Colleague Faked Monkey Data, U.S. Says', *Science*, 315 (5810) (19 January 2007), 317 <http://dx.doi.org/10.1126/science.315.5810.317a>

[42] For further explanation of these examples and others refer to: Larry D. Claxton, 'Scientific authorship: Part 1. A window into scientific fraud?' *Mutation research/Review in Mutation Research*, 589 (1) (2005), 17-30 <http://dx.doi.org/10.1016/j.mrrev.2004.07.003>

[43] Adam Marcus and Ivan Oransky, 'Retraction Watch Blog', *Retraction Watch Website* <http://retractionwatch.wordpress.com/> [accessed 9 August 2015]; Ivan Oransky, 'Embargo Watch Blog', *Embargo Watch Website* <https://embargowatch.wordpress.com/> [accessed 9 August 2015]; Debora Weber-Wulff, 'Copy, Shake and Paste: A Blog about Plagiarism and Scientific Misconduct', *Copy, Shake and Paste Website* <http://copy-shake-paste.blogspot.co.uk/> [accessed 9 August 2015].

[44] Daniele Fanelli, 'Why Growing Retractions Are (Mostly) a Good Sign', *PLoS Medicine*, 10 (12) (2013) e1001563 <http://dx.doi.org/10.1371/journal.pmed.1001563>; Daniele Fanelli, 'Redefine misconduct as distorted reporting', *Nature*, 494 (149) (14 February 2013) <http://dx.doi.org/10.1038/494149a>; R. Grant Steen, Arturo Casadevall, and Ferric C. Fang, 'Why Has the Number of Scientific Retractions Increased?' *PLoS One*, 8 (7) (2013) <http://dx.doi.org/10.1371/journal.pone.0068397>; Carl Zimmer, 'A Sharp Rise in Retractions Prompts Calls for Reform', *New York Times*, 16 April 2012 <http://www.nytimes.com/2012/04/17/science/rise-in-scientific-journal-retractions-prompts-calls-for-reform.html?_r=0> [accessed 9 August 2015].

practical experimentation; and, (3) the ethical violations were not made apparent through the institutional and peer-review processes.

To catch out peer reviewers and the academic community some articles containing unreliable, weak and unfit for purpose academic research data have been precisely produced. An example of this is Philip Davis who submitted a fake, computer-generated knowledge claim that was published by *The Open Information Science Journal*,[45] and reported in *Nature* News Online:

> Davis says he decided to submit the fake manuscript after receiving several unsolicited invitations by e-mail to submit papers to open-access journals published by Bentham under the author-pays-for-publication model. He wanted to test if the publisher would "accept a completely nonsensical manuscript if the authors were willing to pay".[46]

Spoof knowledge claims are based on fabricated academic research data, but they do not have the ill-intention to defraud research users. They seek to test the system that assures the reliability of academic research data for research users. Therefore, any model that assures an academic research datum as reliable, robust and fit for purpose needs to be able to cope with *bone fide* spoof claims. Spoof claims illustrate also how unreliable academic research data can make it successfully through quality assurance to research users. The failures derive from inadequate peer review and overview by journal editors.

### 2.2.3 Secondary research questions: unreliable data – areas for concern

The major problem that the case of Hwang and others revealed is the limited access to and scrutiny of the underlying academic research data, as Lawrence B. Ebert summarises:[47]

> The failure of editors and referees of the journal Science to detect the fraud in manuscripts of Woo Suk Hwang prior to publication, and the widespread acceptance of the work after publication, illustrates some difficulties in relying on peer review to authenticate the validity of scientific work. Neither journals nor the USPTO [The United States Patent and Trademark Office] have the

---

[45] Jessica Shepherd, 'Editor quits after journal accepts bogus science article', *Guardian*, 18 June 2009 <http://www.guardian.co.uk/education/2009/jun/18/science-editor-resigns-hoax-article> [accessed 9 August 2015].

[46] Natasha Gilbert, 'Editor will quit over hoax paper: Computer-generated manuscript accepted for publication in open-access journal', *Nature News*, 15 June 2009 <http://dx.doi.org/10.1038/news.2009.571>

[47] Lawrence B. Ebert, 'Lessons to be Learned from the Hwang Matter: Analyzing Innovation the Right Way', *Patent and Trademark Office Society*, 88 (2006), 239-255 (p. 255).

> resources to perform experiments or even to rigorously review the authenticity of data. Under existing review protocols, fraudulent science or even bad science can pass review.[48]

While self-evident, publication is only as good as the people overseeing the processes. *Science* conducted a report into the publication of the two articles based on fraudulent claims.[49] The Editor-in-Chief of *Science*, Donald Kennedy, maintains that *Science* did follow procedures with regard to their publication. However, the report proposes that a risk-assessment template should be utilised for potential high-risk papers.[50] Donald Kennedy states:

> We are also considering the kinds of special attention that might be given to these high-risk papers. These might include higher standards for including primary data, demands for clearer specification of the roles of all authors, and more intensive evaluation of the treatment of digital images.[51]

The majority of academic journal publishers are not data archives; therefore a distinction needs to be drawn between *Science* as a guarantor of research quality rather than a custodian of the underlying data.[52]

Despite the ease of making underlying data digitally accessible, these were not fully released to *Science* or the wider research community. This limited accessibility meant that the expert peer-reviewers were without means to fully assess the reliability of these articles. While repeat experiments are not always an option, simple checks can be made regarding data generated within cloning experiments, such as DNA testing the tissue samples from the procedure and the verification of image consistency.[53] For instance, on 4 December 2005, Hwang informed *Science* that his 2005 article contained duplicate images, an early warning sign of potential misconduct.[54] Moreover, one of the

---

[48] Ebert, (p. 255).

[49] 'Supporting Online Material for Responding to Fraud', *Science*, 314 (5804) (1 December 2006 ) <http://dx.doi.org/10.1126/science.1137840>

[50] Donald Kennedy, 'Editorial: Responding to Fraud', *Science*, 314 (5804) (1 December 2006 ) 1353 <http://dx.doi.org/10.1126/science.1137840>

[51] *Ibid.*

[52] For further information refer to: Elizabeth Wager and Sabine Kleinert, 'Cooperation between research institutions and journals on research integrity cases', (on behalf of COPE Council) (March 2012), *Committee on Publication Ethics (COPE) Website* <http://publicationethics.org/files/Research_institutions_guidelines_final.pdf> [accessed 9 August 2015].

[53] 'Editorial: Standards for papers on cloning', *Nature*, 432 (243) (19 January 2006) <http://dx.doi.org/10.1038/439243a>

[54] Gretchen Vogel, 'Landmark Paper Has an Image Problem', *Science*, 310 (5754) (9 December 2005) 1595 <http://dx.doi.org/10.1126/science.310.5754.1595>

co-authors, Sun Jong Kim, was alleged to have destroyed some of these data.[55] In consequence, this alleged data destruction potentially threatens the traceability of this research and its validation by independent parties. There are also questions over why the results of these academic misconduct investigations were not made public.[56]

#### 2.2.3.1 Diversity of types of academic research data from multiple originators, contexts and sources

Academic publishers are processing larger quantities of research submissions from a wider variety of sub-disciplines in a digital age, as the research community continues to grow and/or is facing audit such as the UK Research Excellence Framework (REF). Moreover, digital publication processes have facilitated much quicker dissemination rates. For these reasons, it has become more challenging for publishers to ensure that all articles submitted as part of this knowledge explosion are verified to a high standard. In consequence, this significant area for concern raised by the case of Hwang and others leads to the following secondary research question to be addressed in Chapter 4: (Question 3) how should diverse types of academic research data from multiple originators, contexts and sources be safeguarded for a wide set of research users?

#### 2.2.3.2 Multi-author research

Unethical authorship was a key area for concern in the case of Hwang and others.[57] The two articles published by Hwang and others involve a significant number of co-authors, which is usual for many scientific disciplines where research is highly collaborative.[58] However, the extent in which each co-author contributed to these papers, and who the genuine authors were, is unknown.[59]

---

[55] Wohn and Normile.

[56] M. A. G. van der Heyden, T. van de ven and T. Opthof, 'Fraud and misconduct in science: the stem cell seduction', *Netherlands Heart Journal*, 17 (1) (2009), 25-29 (p. 25) <http://ukpmc.ac.uk/articles/PMC2626656/> [accessed 9 August 2015].

[57] COPE offers a number of best practice documents to handle issues such as unethical authorship, refer to: Code of Conduct and Best Practice Guidelines for Journal Editors', (2011) *Committee on Publication Ethics (COPE) Website* <http://publicationethics.org/files/Code_of_conduct_for_journal_editors.pdf> [accessed 9 August 2015].

[58] This is in contrast to the humanities where it is common for a journal article to have one author. The individual author's part is usually clearly visible in multi-authored work in the humanities – there is also a lower amount of co-authoring. Therefore, multi-authorship is more problematic within the sciences.

[59] Ebert, (p. 255).

Following common scientific practices, the name of the principal author appears first in the list of co-authors, which in this case was Woo Suk Hwang.[60] As the highly-regarded principal investigator, his role was to oversee all aspects of research and conduct. Since he later claimed that he was deceived by his junior researchers this oversight and the individual contributions of each researcher are both in doubt. Moreover, one of the co-authors, Gerald Schatten, attempted to remove his name from the paper; although it is very likely that as the co-author he would have completed a copyright declaration. It was later confirmed by a University of Pittsburgh review panel that Schatten had not taken part in the practical experimentation and therefore was not in the position to be named as a senior co-author.[61] By neglecting to verify the role of each co-author, *Science* 'removed one of the barriers to fraud'.[62]

Continuity is a problem for laboratory research teams, as researchers frequently leave and join, which impacts on the integrity and authority of co-authorship. This significant area for concern (raised by the case of Hwang and others) gives rise to further secondary research question to be explored in Chapter 5: (Question 4) how should academic research data generated as part of collaborative research be treated in order to balance a range of difficult communal and continuity problems?

---

[60] Woo Suk Hwang and others (2004); the names of the co-authors as they appear on the 2004 article as follows: Woo Suk Hwang, Young June Ryu, Jong Hyuk Park, Eul Soon Park, Eu Gene Lee, Ja Min Koo, Hyun Yong Jeon, Byeong Chun Lee, Sung Keun Kang, Sun Jong Kim, Curie Ahn, Jung Hye Hwang, Ky Young Park, Jose B. Cibelli, and Shin Yong Moon. Woo Suk Hwang and others, (2005);

The names of the co-authors as they appear on the 2005 article: Woo Suk Hwang, Sung Il Roh, Byeong Chun Lee, Sung Keun Kang, Dae Kee Kwon, Sue Kim, Sun Jong Kim, Sun Woo Park, Hee Sun Kwon, Chang Kyu Lee, Jung Bok Lee, Jin Mee Kim, Curie Ahn, Sun Ha Paek, Sang Sik Chang, Jung Jin Koo, Hyun Soo Yoon, Jung Hye Hwang, Youn Young Hwang, Ye Soo Park, Sun Kyung Oh, Hee Sun Kim, Jong Hyuk Park, Shin Yong Moon, and Gerald Schatten.

[61] For further information about (multi-)authorship issues refer to: Daniel K. Sokol, The dilemma of authorship', *British Medical Journal*, 336 (2008) <http://dx.doi.org/10.1136/bmj.39500.620174.94>; Philip Greenland and Phil B. Fontanarosa, 'Editorial: Ending Honorary Authorship', *Science* 337 (6098) (31 August 2012), 1019 <http://dx.doi.org/10.1126/science.1224988>; Ana Marušić, Lana Bošnjak, and Ana Jerončić, 'A Systematic Review of Research on the Meaning, Ethics and Practices of Authorship across Scholarly Disciplines', *PLoS One*, 6 (9) (8 September 2011) e23477<http://dx.doi.org/10.1371/journal.pone.0023477>

[62] Ebert, (p. 255).

#### 2.2.3.3 Unethical practices

As this research was conducted at a well-established higher education institution, Seoul National University, it is rational to assume that there would be vigorous institutional checks of compliance with both legal and ethical stipulations. Particularly as bioethics is a controversial area, it is essential that higher education institutions enforce a code of good research practice. Despite this, Hwang and others were in serious breach of ethical codes of conduct to the extent the Seoul Central District Court held that Hwang was guilty of purchasing human eggs in breach of bioethics law. It could be argued therefore that due to its sensitive nature, these data could not be made available for public scrutiny without some form of redaction. This significant area for concern (raised by the case of Hwang and others) gives rise to the final secondary research question to be explored in Chapter 6: (Question 5) how should maintenance of and access to sensitive and personalised data be treated in order to balance a range of difficult problems with permissions, data protection and confidentiality?

### 2.2.4 More stem cell research controversy

Regrettably, controversy surrounding stem cell research continues with the recent and high profile retractions of two articles published by Haruko Obokata and others in *Nature*.[63] A retraction statement was published by the authors on 2 July 2014; which for James Gallagher, the health editor for *BBC News* website, 'brings back memories of the false claims' produced by Hwang and others eight years earlier.[64] Thus, this shows that the calls made during 2006 to strengthen scrutiny of underlying and supporting academic research data had still to be fully realised in 2014. In consequence, the very issues this thesis addresses remain highly significant both for modelling best practice now and in the future.

---

[63] Haruko Obokata and others, 'Retraction: Bidirectional developmental potential in reprogrammed cells with acquired pluripotency', *Nature*, 511 (7507) (2014), 112 <http://dx.doi.org/10.1038/nature13599>; 'STAP retracted: Two retractions highlight long-standing issues of trust and sloppiness that must be addressed', *Nature*, 511 (7507), (2 July 2014), 5-6 <http://dx.doi.org/10.1038/511005b>; Ian Sample, 'Stem cell scientist Haruko Obokata found guilty of misconduct', *Guardian*, 1 April 2014 <http://www.theguardian.com/science/2014/apr/01/stem-cell-scientist-haruko-obokata-guilty-misconduct-committee> [accessed 9 August 2015].

[64] James Gallagher, 'Japanese stem-cell 'breakthrough' findings retracted', *BBC News*, 2 July 2014 <http://www.bbc.co.uk/news/health-28124749> [accessed 9 August 2015].

## 2.3 Print to ePrint

Throughout the course of human history knowledge transfer has been beleaguered by issues of accessibility, trust and authorisation. However, there is often the tendency to overlook or dismiss the historical and contextual development of knowledge transfer, as Lauren Rabinovitz and Abraham Geil state:[65]

> Discussions about digital culture assume that new computerized technologies provide such fundamental rupture from the past that there are no continuities or, worse, that they willfully obliterate the past in creating new models. Such ahistoricism is problematic because it tends to reproduce at the level of scholarship what is one of the hallmarks of digital culture – its rhetoric of newness. It is painfully obvious that this is neither the first technological revolution in human history nor an event independent from its cultural heritages and historical roots, and so a rhetoric of newness is at best a myopic one.[66]

The transfer of knowledge is a human activity that has occurred across oral, manuscript/scribal, print and digital ages.[67] The Web is not the first technological development that has significantly contributed to its reformation. In the 1400s, the emergence of the printing press in Europe (its European originators include Gutenberg, Manutius and Caxton) sought to challenge the professional scribe of the manuscript age.[68] As print culture matured, the printing press became a system of mass production

---

[65] Lauren Rabinovitz and Abraham Geil, 'Introduction' in *Memory Bytes: History, Technology, and Digital Culture*, ed. by Lauren Rabinovitz and Abraham Geil (Durham NC: Duke University Press, 2004), pp.1-2. Google eBook.

[66] Rabinovitz and Geil, pp.1-2.

[67] For further information on the transition between oral, manuscript, print and digital cultures refer to: Cushla Kapitzke, 'Ceremony and cybrary: Digital libraries and the dialectic of place and space', *Social Alternatives*, 20 (1) (2001), 33-40 <http://eprints.qut.edu.au/43995/1/Kapitzke_cybrary_Social_Alternatives.pdf> [accessed 9 August 2015].ePrint.

[68] For more information about print culture and the printing press revolution refer to: Elizabeth L. Eisenstein, *The Printing Revolution in early Modern Europe* 2nd edn (New York, NY: Cambridge University Press, 2005). Google eBook; Adrian Johns, *The Nature of the Book: Print and Knowledge in the Making* (Chicago: The University of Chicago Press, 1998). Google eBook.

However, it must be noted that, in the eleventh century, China already had moveable type. To what extent therefore the Gutenberg press is indebted to this different form of moveable type is to be speculated, refer to: Christopher Cullen, 'Reflections on the Transmission and Transformation of Technologies: Agriculture, Printing and Gunpowder between East and West' in *Science Between Europe and Asia: Historical studies on the transmission, adoption and adaption of knowledge*, ed. by Feza Günergun and Dhruv Raina (London: Springer Dordrecht Heidelberg, 2011), pp. 13-26. Google eBook.

and distribution for data, information and knowledge.[69] However, this greater access was not only a simple technological change.[70] Print was only brought to the masses by cheaper printing technologies and a rise in literacy. Much of this change was witnessed during the 1700s, as Carey McIntosh states:

> At the beginning of the century, there was no adequate copyright, only four towns could boast of printing presses (London, Oxford, Cambridge, York), and only a few score people were making a living in the world of publishing, journalism and advertising. By the end of the century there were words for sale in every village of the nation.[71]

It is outside the scope of the thesis to offer more than this brief overview of the development of the print age. Moreover, the esteemed works of both Elizabeth L. Eisenstein and Adrian Johns already critically assess the impact of the print age in great detail, and should be referred to for further information.[72] Eisenstein focuses on the emergence of the print culture in Europe (including the republic of letters) within its wider socio-cultural context. In parallel, Johns examines print and culture in the making, and the socio-cultural and technical processes that shaped its development.

In consequence, it is of paramount importance that in order to better understand how to model best practice, this thesis first addresses and embraces the historical and contextual development of knowledge transfer. As a result, this section explores what lessons can be learnt from past practices, and determines whether the accessibility, authorisation, trust and traceability issues around data re-usage are largely a manifestation of long-standing problems in a digital age.

---

[69] *The Sage Glossary of the Social and Behavioral Sciences*, ed. by Larry E. Sullivan (California: SAGE Publications, 2009), p. 402. Google eBook.

[70] It is outside the remit of this thesis to examine sociological theories about the relationship between technology and society. However, this thesis rejects technological determinist and technological instrumentalist arguments. For more information refer to: Graeme Salter, 'Factors Affecting the Adoption of Educational Technology' in *Encyclopedia of Distance Learning*, ed. by Caroline Howard, (London: Idea Group, 2005), pp. 922-929 (p. 923).

[71] Carey McIntosh, *The Evolution of English Prose, 1700-1800: Style, Politeness, and Print Culture* (Cambridge, Cambridge University Press, 1998), p. 5.

[72] Eisenstein; Johns.

### 2.3.1 Accessible data

No other technological development has caused such significant change to knowledge transfer, since the printing press, as the emergence of the Web.[73] The Web has the potential to facilitate instant and on-demand access to data on a global scale. However, digital accessibility is not pre-given, because not all data are openly released. Many individuals do not share data often because circumstances prevent the open release of data, such as where datasets contain particularly personal and/or sensitive materials. Moreover not everyone has access to the Web.[74]

Pleas for more accessible data and information have roots in the print era. As early as the 1600s, the concept of the republic of letters was voiced by Thomas Bodley to facilitate a community for knowledge transfer.[75] Paul Keen further explains the scope of the republic of letters:

> Literature, or the republic of letters, as it was often referred to, was celebrated by the advocates of this vision as a basis of a communicative process in which all rational individuals could have their say, and in which an increasingly enlightened reading public would be able to judge the merit of different arguments for themselves. […] The hopes and anxieties generated by this communicative ideal have strong parallels with responses to the 'information revolution' in our own age. Although rooted in the printing press rather than computers (the Internet or World Wide Web, Electronic publishing), it was similarly discussed in terms of empowerment, rationalization, and inevitably, alienation.[76]

The republic of letters appear to have re-manifested in the digital age with the emergence of various open movements – such as open access, open licensing, open source, and open data. The latter category specifically aims to promote the public release of data where possible.

The digital open access movement was an important precursor to the digital open data movement, which aims to openly release data from a number of key sectors,

---

[73] Ingrid Silver and Helen Anderson, 'Gutenberg odyssey - the advent of e-books and some implications for the world of publishing', *Entertainment Law Review*, 21 (6) (2010), 225-228 (p. 225); Yoshiyuki Tamura, 'Rethinking copyright institution for the digital age', *WIPO Journal*, 1 (2009), 63-74 (p. 66).

[74] For more information on the 'digital divide' refer to: Haochen Sun, 'Copyright law under siege: an inquiry into the legitimacy of copyright protection in the context of the global digital divide', *International Review of Intellectual Property and Competition Law*, 36 (2) (2005), 192–213.

[75] G.R, Evans, 'Academic Libraries and the Law: What Legal Protections Guarantee the Survival of Britain's Academic Library Collections?' *Education Law*, 4 (2008), 248 (p. 248).

[76] Paul Keen, *The crisis of literature in the 1790s: print culture and the public sphere* (Cambridge: Cambridge University Press, 1999), p. 4.

including government and academic research. The digital open access movement was ignited through three conferences in the early 2000s.[77] Open access aims to make all scholarly works available (where possible) without charge at the point of use.[78] Therefore, it intends to remove subscription fee barriers on access by end users.[79] There are two main routes to open access: the golden route and the green route. The golden route involves an author, funding body and/or institution paying publishing fees (known as author pays) to a journal instead of the traditional subscription model where libraries would pay to access the work. Otherwise another golden route to open access is publication within an open access journal. The green route requires authors to self-archive their work and authorise complete unrestricted access to that work (subject to authorisation, where necessary, of the journal publisher).

Some of the key events in the history of the open access movement were originally recorded by Peter Suber.[80] Since February 2009, this record – the Timeline of the Open Access Movement – was copied to the Open Access Directory and is open for community editing.[81] For instance, Peter Murray-Rust and Henry Rzepa have both made significant contributions to the open access movements within chemistry largely through the creation of Chemical Mark-up Language (CML).[82] Murray-Rust and Rzepa

---

[77] Dinusha Kishani Mendis, *Universities and Copyright Collecting Societies* (Cambridge: Cambridge University Press, 2009), pp. 205-6; 'View signatures', *Budapest Open Access Initiative Website* <http://www.budapestopenaccessinitiative.org/list_signatures> [accessed 9 August 2015]; 'Bethesda Statement on Open Access Publishing', *Bethesda Open Access Website* <http://www.earlham.edu/~peters/fos/bethesda.htm#summary>[ accessed 9 August 2015]; 'Open Access to Knowledge in the Sciences and Humanities', Location: Harnack House of the Max Planck Society, Berlin-Dahlem, Germany (20-22 October 2003), *Open Access Max-Planck-Gesellschaft Website* <http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/> [accessed 9 August 2015].

[78] Open access issues are the focus of the thesis author's Web Science MSc dissertation: Laura German, 'To what extent is open access changing scholarly communication within the humanities.' Unpublished MSc (Web Science) Dissertation, University of Southampton, submitted on 24 September 2010.

[79] For more information about green and golden routes to open access refer to: Steven Harnad and others, 'The Access/Impact Problem and the Green and Gold Roads to Open Access', *Serials Review*, 30 (4) (2004) <http://dx.doi.org/10.1016/j.serrev.2004.09.013>; 'Better access to scientific articles on EU-funded research: online pilot project', *European Union Focus* 240 (2008), 24-25 (p. 24).

[80] Peter Suber, 'Timeline of the Open Access Movement', last revised 9 February 2009 <http://legacy.earlham.edu/~peters/fos/timeline.htm> [accessed 9 August 2015].

[81] 'Timeline', *Open Access Directory Website* <http://oad.simmons.edu/oadwiki/Timeline> [accessed 9 August 2015].

[82] Refer to the following webpage for a number of useful articles by Peter Murray-Rust, Henry Rzepa and others: 'Chemical Markup Language – publications', *Chemical Mark-up Language – CML Website* <http://www.xml-cml.org/documentation/biblio.html> [accessed 9 August 2015].

created CML (in 1995) as the XML standard for chemistry, which provides a common and open format for data sharing within the discipline.[83] Murray-Rust also co-authored the Panton Principles for Open Data in Science.[84]

April 2013 not only observed the twentieth anniversary of the Web, but witnessed the enactment of the Research Councils UK's Policy on Open Access.[85] This policy mainly focuses on academic research publication rather than academic research data, but it does outline academic research data sharing expectations:

> **3.3 Acknowledgement of funding sources and underlying research material** [/…/] (ii) As part of supporting the drive for openness and transparency in research, and to ensure that researchers think about data access issues, the policy requires all research papers, if applicable, to include a statement on how underlying research materials, such as data, samples or models, can be accessed. However, the policy does not require that the data must be made open. If there are considered to be compelling reasons to protect access to the data, for example commercial confidentiality or legitimate sensitivities around data derived from potentially identifiable human participants, these should be included in the statement.[86] [RCUK's bold emphasis and underline.]

This RCUK policy compels all researchers within UK higher education institutions to provide a statement where applicable to outline where data are made accessible. Academic research data re-usage is not a necessity however.

Although only recently identified by the term open data, the sustainable management and open release of data is not a new concept. Data centres for the safeguard of (academic research) data exist across print and e-print eras. For instance, on a national level, the UK Data Archive 'is curator of the largest collection of digital data in the social sciences and humanities in the United Kingdom.'[87] However, this archive was founded prior to the emergence of the Web in 1967.[88]

---

[83] *Chemical Mark-up Language (CML) Website* <http://www.xml-cml.org/> [accessed 9 August 2015].

[84] Peter Murray-Rust, Cameron Neylon, Rufus Pollock and John Wilbanks, 'Panton Principles for Open Data in Science', 19 February 2010, *Panton Principles Website* <http://pantonprinciples.org/> [accessed 9 August 2015].

[85] 'RCUK announces block grants for universities to aid drives to open access to research outputs', *Research Councils UK' (RCUK) Website,* 8 November 2012 <http://www.rcuk.ac.uk/media/news/121108/> [accessed 9 August 2015].

[86] RCUK, 'RCUK Policy on Open Access and Supporting Guidance', *Research Councils UK (RCUK) Website* (p. 4) <http://www.rcuk.ac.uk/documents/documents/RCUKOpenAccessPolicy.pdf> [accessed 9 August 2015].

[87] *The UK Data Archive Website* <http://www.data-archive.ac.uk/about/archive> [accessed 9 August 2015].

[88] *Ibid.*

An increased number of data management platforms (made possible by the Web) are able to capture and facilitate more data as open data in a digital age. For example, on an international level, figshare facilitates self-publication by providing: 'a repository where users can make all of their research outputs available in a citable, shareable and discoverable manner.'[89] There are also a number of institutional data repositories, such as at the University of Edinburgh.[90]

There is a paradox between data scarcity (where data are not fully released for scrutiny) and data overload (where self-publication has significantly contributed to a colossal number of digitally accessible data). For instance, many academic journals do not check underlying data or ask for copies of datasets, such in the case of Hwang and others. However, this data scarcity/overload paradox is not a new digital phenomenon. In 1945, Vannevar Bush – a hypertext pioneer – describes the problem of an overload or deluge of data:

> There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers—conclusions which he cannot find time to grasp, much less to remember, as they appear. Yet specialization becomes increasingly necessary for progress, and the effort to bridge between disciplines is correspondingly superficial. [91]

While the digital data overload problem is a manifestation of the information deluge faced by a mature print culture, the continued global expansion of the higher education sector has significantly contributed to a growing amount of academic research in a digital age. In consequence, it is essential that research users are able to effectively navigate through available academic research data to source the highest quality datasets for re-usage.

In summary, the various digital open movements of the e-print era are firmly rooted in the republic of letters. Greater technological capabilities may have made it easier to openly release and access data on a global scale, but open data is not guaranteed. Data accessibility is re-visited in Chapters 4-6, as all cases studies aim to

---

[89] 'About', *figshare Website* <http://figshare.com/about> [accessed 9 August 2015].

[90] 'Edinburgh DataShare', *University of Edinburgh Website* <http://www.ed.ac.uk/schools-departments/information-services/research-support/data-library/data-repository> [accessed 9 August 2015].

[91] Vannevar Bush, 'As we may think', *Atlantic Monthly*, (1 July 1945) <http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/> [accessed 9 August 2015].

encourage and make data openly accessible where possible. For instance, in the case of geo-spatial data, European legislation – the Infrastructure for Spatial Information in Europe (INSPIRE) Directive (2007/2/EC) – is facilitating the sharing of spatial data between member states through common standards, such as metadata and data specifications (see Chapter 4 for full information).[92]

### 2.3.2 Trust issues

Accessibility is only one factor ensuring that academic research data are re-usable. The trustworthiness of knowledge claims has been the focus for peer-review from print to e-print eras. The user-community must have confidence in the reliability of data in order to re-use particular datasets or articles which are built on certain underlying datasets. Without trust, knowledge transfer is of little value.

As early as 1997, Tim Berners-Lee was contemplating ways in which trust could be established on the Web with the theoretical "Oh, yeah?" button.[93] It was envisaged that such a button could be pressed on any webpage by a user to verify the information contained therein. A user query would then receive one of three responses: (1) a list of assumptions on which the information should be trusted; (2) an error message stating that such authentication was not possible; or, (3) that no assumptions could be found.

Trust and trustworthiness remains crucial to knowledge transfer. Given the global expansion of the academic community and subsequent increase in academic research production, trust can sometimes be more difficult to establish due to an increasing volume of academic networks. Katherine Gross and Gary Mittelbach state:

> Today the sheer size of the scientific community and its global distribution make it impossible to know many of science's practitioners personally. Moreover, the integrity of science cannot simply rely on the trustworthiness of researchers for two important reasons. First scientists are human. As such they are subject to the failings of ambition, vanity and greed, as well as the external influences of culture and politics. Second, even the best-intentioned scientists can get so caught up in their own worldview that they cling to their favorite hypothesis or refuse to accept observations that are inconsistent with their view.[94]

---

[92] *Infrastructure for Spatial Information in Europe (INSPIRE) Directive Website* <http://inspire.ec.europa.eu/> [accessed 9 August 2015].

[93] Tim Berners-Lee, 'The "Oh yeah?" button'. W3C: Design Issues, Status: personal view, 6 February 1997, *3C Website* <http://www.w3.org/DesignIssues/UI.html#OhYeah> [accessed 9 August 2015].

[94] Katherine L. Gross and Gary G. Mittelbach, 'What Maintains the Integrity of Science: An Essay for Nonscientists', *Emory Law Journal*, 58 (2008-2009), 341-356 (p. 344.)

Codes of conduct have grown in number and size to help ensure best practice within academic networks. This includes: the four Panton Principles: Principles for Open Data in Science (2010) and Tim Berners-Lee's publication of a star rating system for linked open data on the Web (2010).[95] Independent organisations have also been founded to offer best practice guidance on data re-usage. Founded in 1997, the Committee on Publication Ethics (COPE) is one such organisation, which provides guidance to editors and publishers on research misconduct.[96] Founded in 2004, the Digital Curation Centre (DCC) is another important organisation:

> The Digital Curation Centre provides expert advice and practical help to anyone in UK higher education and research wanting to store, manage, protect and share digital research data. [/] The DCC provides access to a range of resources including our popular How-to Guides, case studies and online services. Our training programmes aim to equip researchers and data custodians with the skills they need to manage and share data effectively.[97]

In addition to the codes stipulated by funding councils and higher education institutions, a number of these protocols have been developed by a range of organisations over the past ten years, such as the Open Knowledge Foundation, the Knowledge Exchange, the Open Data Institute, and the Research Data Alliance.[98]

Soft approaches, such as (voluntary) codes of conduct, are crucial for raising-awareness of best practice. Hard approaches (e.g. the processing of personal and sensitive academic research data will be subject to the Data Protection Act 1998) are also vital to provide mandatory minimum standards for compliance, mechanisms for enforcement and damages. However, such moral and legal codes are only as robust as the people adhering to them. Trust is predicated not only on the quality and ethical standing of an academic research dataset, but that a dataset has been authorised for re-usage. Providing resources of trustworthy academic research data is a key feature

---

[95] Murray-Rust and others, 'Panton Principles for Open Data in Science'; Tim Berners-Lee, 'Linked data', W3C: Design Issues, Status: personal view, 18 June 2009, *3C Website* <http://www.w3.org/DesignIssues/LinkedData.html> [accessed 9 August 2015].

[96] 'About COPE', *Committee on Publication Ethics (COPE) Website* <http://publicationethics.org/about> [accessed 9 August 2015].

[97] 'About the DCC', *Digital Curation Centre (DCC) Website* <http://www.dcc.ac.uk/about-us> [accessed 9 August 2015].

[98] *Open Knowledge Foundation Website* <https://okfn.org/> [accessed 9 August 2015]; *Knowledge Exchange Website* <http://www.knowledge-exchange.info/> [accessed 9 August 2015]; *Open Data Institute Website* <http://opendatainstitute.org/> [accessed 9 August 2015]; *Research Data Alliance Website* <https://rd-alliance.org/> [accessed 9 August 2015].

throughout Chapters 4-6. Despite managing a diverse range of contextual issues, each data platform strives to provide high quality datasets.

### 2.3.3 Authorisation issues

Data originators, custodians and research users all need to be both ethically and legally permitted to collect, store, manage, disseminate and re-use academic research data. Therefore, non-rights holders should be appropriately licensed to carry out acts which go beyond exemptions permitted by law; e.g. to use a greater part of a copyright work than permitted by fair dealing. Therefore intellectual property law – particularly the areas of copyright, the *sui generis* database right and licensing – plays a significant role in the re-usage of academic research data.[99]

The legal control exerted by rights holders over their intellectual works is not a new phenomenon. The commodification and control of knowledge transfer are long established concepts in both pre-digital and digital ages. For instance, in second century ancient Rome, trade agreements between publishers were established to prevent publishers copying each other's books.[100] Printing privileges were also granted in ancient China, although Western commentators often focus on the country's historical lack of copyright law.[101] This example of an ancient Chinese printing privilege is very similar to those granted by the Crown which led up to the Statute of Anne 1710 – 'For the encouragement of Learning and for securing the property of copies of the books to the rightful owners thereof.'[102] This Statute is widely-regarded as the foundation for copyright law in the UK.

---

[99] For a comprehensive overview of intellectual property law refer to the information provided by the Intellectual Property Office via the *UK Government Website* <https://www.gov.uk/government/organisations/intellectual-property-office/> [accessed 9 August 2015].

[100] Matt Jackson, 'From Private to Public: Reexamining the Technological Basis for Copyright', *Journal of Communication*, 53 (2) (2002), 416-433 (p. 419) <http://dx.doi.org/10.1111/j.1460-2466.2002.tb02553.x> ; Matt Jackson cites: George Haven Putnam, *Authors and Their Public in Ancient Times* (New York: Putnam, 1893); H. L. Pinner, *The World of Books in Classical Antiquity* (Leiden, Netherlands: A.W. Sijthoff, 1958), p. 38.

[101] Peter Ganea, 'C. Copyright: I Historical Overview' in *Intellectual Property Law in China*, ed. by Peter Ganea, Thomas Pattloch and Christopher Heath (The Hague, the Netherlands: Kluwer Law International, 2005), pp.205-213 (pp. 206-207). Google eBook.

[102] For a digital version of the Statute of Anne 1710 refer to: Lionel Bently and Martin Kretschmer (eds.) *Primary Sources on Copyright (1450-1900) Website* <www.copyrighthistory.org> [accessed 9 August 2015]; 'The Statute of Anne; April 10, 1710', *The Avalon Project Lillian Goldman Law Library, Yale Law*

Henry Self states that: 'Western ideals of intellectual property rights are inextricably linked to fundamental concepts of ownership, exclusion and most importantly – capitalism.'[103] However, for researchers within higher education the commodification of research is not a primary driver for sharing academic research data.[104] The principal benefits lie in non-financial reward such as building an academic reputation and the possibility of advancing human knowledge. The UK copyright law framework is based on an economic system, and thus is not always at ease with academic research, as Charlotte Waelde and Hector MacQueen contend:

> much of the reform of copyright law which has occurred since the mid-1990s has been driven by the concerns of what we call the "entertainment industry" […/…] Relatively little has been heard as yet about the impact the policies will have upon the interests of education and research and the sectors, private and public, which support and provide for these interests. There have, however, been signs of stress in the relationship between copyright law and higher education teaching and research in the United Kingdom. [105]

---

*School, Yale University Website* <http://avalon.law.yale.edu/18th_century/anne_1710.asp> [accessed 9 August 2015].

For further information on the history of copyright refer to: Augustine Birrell, *Seven Lectures on the Law and History of Copyright in Books* (London: Cassell and Company, 1899) <http://archive.org/stream/cu31924029522061#page/n5/mode/2up> [accessed 9 August 2015]; Ronan Deazley, *On the Origin of the Right to Copy: Charting the Movement of Copyright Law in Eighteenth Century Britain (1695-1775)* (Oxford: Hart Publishing, 2004). Google eBook; Ronan Deazley, *Rethinking Copyright: History, Theory, Language* (Cheltenham: Edward Elgar, 2006). Google eBook; *Privilege and Property: Essays on the History of Copyright*, ed. by Ronan Deazley, Martin Kretschmer, and Lionel Bently (Cambridge: Open Book Publishers, 2010). Google eBook; Joseph Loewenstein, *The Author's Due: Printing and the Prehistory of Copyright* (Chicago: The University of Chicago Press, 2002). Google eBook; Mark Rose, *Authors and Owners: The Invention of Copyright* (Cambridge MA: Harvard University Press, 1993). Google eBook; Melissa De Zwart, 'A historical analysis of the birth of fair dealing and fair use: lessons for the digital age', *Intellectual Property Quarterly*, 1 (2007), 60-91.

[103] Henry Self, 'Digital Sampling: A Cultural Perspective', *UCLA Entertainment Law Review*, 9 (2) (2002), 347-359 (p. 357); for further information Self recommends: Marci A. Hamilton, 'The TRIPS Agreement: Imperialistic, Outdated and Overprotective', *Vanderbilt Journal of Transnational Law*, 29 (1996) 616-617; also refer to: Karl-Nikolaus Peifer, 'The return of the commons – copyright history as a helpful source?' *International Review of Intellectual Property and Competition Law*, 39 (6) (2008), 679-688, p. 682.

[104] For a comprehensive overview of the philosophical basis for the justification of copyright law refer to: Horacio M. Spector, 'An outline of a theory justifying intellectual and industrial property rights', *European Intellectual Property Review*, 11 (8) (1989), 270-273.

[105] Charlotte Waelde and Hector MacQueen, 'From entertainment to education: the scope of copyright?' *Intellectual Property Quarterly*, 3 (2004), 259-283 (p. 259).

In this growing myriad of academic research data, it is both impracticable and unrealistic for research users to directly contact the rights holder to ask for permission to re-use a dataset. In consequence, (data) licence agreements are an effective means for rights holders to stipulate which re-usage acts they permit third parties to carry out (in addition to the exemptions granted by law).[106] Rights holders are able to authorise and decide upon the parameters of re-usage pertaining to a particular dataset or set of datasets (e.g. a database). For instance, some rights holders may want to restrict the use of their dataset to non-commercial purposes only.

Licences have therefore become the legislative basis for academic research data re-usage; particularly through the widespread utilisation of open standard licensing (discussed in more detail below).[107] Data originators and other rights holders have the ability to personally control the extent in which their academic research data are re-used. Many individuals within the academic community may opt to select an open standard licence (where terms and conditions are pre-defined by an authoritative third party licence provider). It is increasingly common for data originators to license the re-usage of their academic research data by open standard licences, such as those provided by Creative Commons, rather than reliance on bespoke terms and conditions alone.[108] These licences operate within the existing intellectual property framework and are therefore predicated on the subsistence of copyright and/or a *sui generis* database right.

The Digital Curation Centre (DCC) – an organisation providing expert advice and practical assistance concerning data management to the UK higher education community – offers a useful guide on how to license research data and further provides examples of key open standard licence providers, including: (1) Creative Commons, (2)

---

[106] Although published in 2001 before the emergence of well-known open licensing models, the following article provides a useful overview of what licence agreements contain: Laurence W. Bebbington, 'Managing content: licensing, copyright and privacy issues in managing electronic resources', *Legal Information Management*, 1 (2) (2001), 4-13.

[107] For further information refer to: Tony Simmonds, 'Common knowledge? The rise of Creative Commons licensing', *Legal Information Management*, 10 (3) (2010), 162-165; Mark Fox, Tony Ciro and Nancy Duncan, 'Creative Commons: an alternative, web-based copyright system', *Entertainment Law Review*, 16 (5) (2005), 111-116; Joelle Farchy, 'Are free licences suitable for cultural works?' *European Intellectual Property Review*, 31 (5) (2009), 255-263; Simone Aliprandi, 'Open licensing and databases', *International Free and Open Source Software Law Review*, 4 (1) (2012), 5-18 <http://www.ifosslr.org/ifosslr/article/view/62/116> [accessed 9 August 2015].

[108] *Creative Commons Website* <https://creativecommons.org/> [accessed 9 August 2015].

Open Data Commons, and (3) the Open Government Licence.[109] Licences from these three providers feature in the case studies (see Chapters 4-6 for further information) and are now briefly introduced in turn.

Founded in 2001, Creative Commons is a non-profit organisation that facilitates the sharing and re-usage of copyright works through openly accessible standard licences, which has become a popular choice amongst researchers.[110] In 2009, it was estimated that 350 million works were provided under a Creative Commons licence.[111] Creative Commons currently offers six types of standard licences: Attribution – CC BY; Attribution-NoDerivs – CC BY-ND; Attribution-NonCommercial-ShareAlike – CC BY-NC-SA; Attribution-ShareAlike – CC BY-SA; Attribution-NonCommercial – CC BY-NC; and, Attribution-NonCommercial-NoDerivs – CC BY-NC-ND.[112] Creative Commons licensing can be used to authorise the re-usage of an array of copyright works from images to video recordings on a national and international scale.

The Open Government Licence is the standard licence for UK public sector information and works.[113] This licence enables UK government departments and public sector bodies to publically release (in perpetuity) certain information for re-usage under a simple set of terms and conditions.[114] Personal data are not covered by this licence however.[115] This Open Government Licence is also compatible with the Creative Commons Attribution License and the Open Data Commons Attribution License. [116] The National Archives provides guidance for information providers and users releasing and re-using information subject to the Open Government Licence.[117] The Open

---

[109] 'How to license research data', *Digital Curation Centre (DCC) Website* <http://www.dcc.ac.uk/resources/how-guides/license-research-data> [accessed 9 August 2015].

[110] *Creative Commons Website* <https://creativecommons.org/> [accessed 9 August 2015].

[111] 'History', *Creative Commons (CC) Website* <http://creativecommons.org/about/history> [accessed 9 August 2015].

[112] 'About the licenses', *Creative Commons Website* <https://creativecommons.org/licenses/> [accessed 9 August 2015].

[113] 'Open Government Licence for public sector information', *The National Archives Website* <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/> [accessed 9 August 2015].

[114] 'About the Open Government Licence', *The National Archives Website* <http://www.nationalarchives.gov.uk/information-management/re-using-public-sector-information/licensing-for-re-use/what-ogl-covers/> [accessed 9 August 2015].

[115] 'About the Open Government Licence', *The National Archives Website.*

[116] *Ibid.*

[117] 'About the Open Government Licence', *The National Archives Website.*

Government Licence is therefore an important aspect of the open government data movement by providing a standard and approved method for authorising the release and re-usage of various data types across a multitude of government departments and bodies.

Launched in March 2008, Open Data Commons provides three licences specifically for data: Public Domain Dedication and License (PDDL) – Public Domain for data/databases; Attribution License (ODC-By) – Attribution for data/databases; and, Open Database License (ODC-ODbL) – Attribution Share-Alike for data/databases.[118] Open Data Commons was launched in 2007 by Jordan Hatcher who wrote the first open database licence (PDDL) with Charlotte Waelde (this work was funded by Talis).[119] These licences are important as they have been produced specifically for data re-usage (unlike Creative Commons which can be used for a variety of different copyright works) and are not limited to one sector (i.e. the Open Government Licence).

While many academic research datasets are now licensed for re-usage, a significant grey area is whether copyright subsists within academic research data.[120] The standard for originality in UK copyright law is low, for example copyright has been held to subsist in football coupons.[121] Nevertheless, some academic research data will fail to meet the legal requirement of originality, such as a list of temperature readings. In some instances licensing will therefore be over-encompassing. For example, a researcher may not realise that an academic research dataset does not meet the standard of originality and incorrectly releases a dataset under a data licence agreement.

However, there appears to be a major misconception that a significant number of academic research data will fail to meet the legal requirement of originality. Eric C. Kansa and others argue that the legal conceptualisation of scientific facts and expression

---

[118] 'About', *Open Data Commons Website* <http://opendatacommons.org/about/> [accessed 9 August 2015]; 'Licenses', *Open Data Commons Website* <http://opendatacommons.org/licenses/> [accessed 9 August 2015].

[119] 'About', *Open Data Commons Website* <http://opendatacommons.org/about/> [accessed 9 August 2015].

[120] John M. Carson and Brian C. Leubitz, 'Case Comment: United States: copyright - protection of databases'. *European Intellectual Property Review*, 26 (5) (2004), N74-75.

[121] *Ladbroke (Football) Ltd v. William Hill (Football) Ltd* [1964] 1 WLR 273; Mark Sherwood-Edwards, 'The redundancy of originality', *Entertainment Law Review,* 6 (3) (1995), 94-106; Gary Lea, 'In defence of originality', *Entertainment Law Review*, 7 (1) (1996), 21-26.

is at odds with how scientific data are recorded and presented in practice.[122] A scientist conducting field research may record their academic research data through tables, graphs, photographs, drawings and written narratives – all of which may constitute individual copyright works.[123] Kansa and others assert:

> Thus, the copyright status of much field documentation is likely to be mixed (depending on the specifics of the records involved) and likely open to interpretation. In any case, the threshold for copyrightable originality is very low and the risks of infringement are extremely high, so a typical user must almost always assume that copyright protections pertain, even if data compilations seem very factual.[124]

In consequence, the debate around whether copyright subsists in academic research is a complex matter with no definitive answer. However, for many types of academic research datasets copyright and/or a *sui generis* database right will be in subsistence, and therefore many datasets are licensed for re-use with the UK higher education sector.

Re-usage rights are not the only important part of licensing – attribution is very significant as well.[125] Good citation and referencing forms part of best academic practice not only to prevent plagiarism, but for traceability. Copyright law is predicated on known authors and rights holders. If attribution information is deleted or missing and data originators and right holders cannot be traced, this gives rise to orphan works, as Stef Van Gompel states:[126]

> The process of clearing rights may be obstructed, however, if one or more right owners of a work or other protected subject matter remain unidentifiable or untraceable after a reasonable search has been conducted by a person intending to use this work. This is the so-called problem of "orphan works". [/] Not being

---

[122] Eric C. Kansa, Jason Schultz and Ahrash N. Bissell, 'Protecting traditional knowledge and expanding access to scientific data: juxtaposing intellectual property agendas via a "some rights reserved" model'. *International Journal of Cultural Property*, 12 (3) (2005), 285-314 (, p. 293).

[123] Kansa, Schultz and Bissell, (p. 293).

[124] *Ibid.*

[125] The concept of authorship continues to change however. It has developed over pre-digital and digital ages from the collaborative authorship in oral cultures and the professional scribe in the manuscript age to the rise of the romanticised concept of the genius after the Statute of Anne, refer to the following literature for more information: Laura L. Mendelson, Privatizing Knowledge: The Demise of Fair Use and the Public University', *Albany Law Journal of Science & Technology*, 13 (2) (2003), 593-612; Craig Baehr and Bob Schaller, *Writing for the Internet: A Guide to Real Communication in Virtual Space* (California: Greenwood Press, 2010), p. 9. Google eBook.

[126] Stef Van Gompel, 'Unlocking the potential of pre-existing content: how to address the issue of orphan works in Europe?' *International Review of Intellectual Property and Competition Law*, 38 (6) (2007), 669-702 (pp. 670-671).

> able to acquire permission from the right owner(s) concerned makes it impossible to reutilise the work legally.[127]

Therefore, maintaining the links between data originators and right holders is essential for sustainable re-usage.[128]

Most intellectual innovation is built on the work of others; a concept that Graham M. Dutfield and Uma Suthersanen describe as 'cumulative creativity'.[129] Attribution operates both through legislative requirement and social norms, as Matteo Migheli and Giovanni B. Ramello state:[130]

> The common purpose of both social norms and laws is to constrain the behaviour of social actors. Social norms emerge spontaneously in human groups and, despite being somewhat informal (they are not promulgated by a legislature, and there is no legal penalty for infringement), can still constrain and regulate a great deal of social interaction. [/…] In some cases social norms are written into and replaced by laws, while in others the two systems can coexist. In particular, social norms are important even when a full body of laws governs society, especially within groups where collective action and reciprocal recognition prevail over a top-down structure.[131]

Four moral rights are conferred by the Copyright, Designs and Patents Act (CDPA) 1988, section 79 to section 89 in the UK: (1) right of paternity; (2) the right to integrity; (3) the right to object to false attribution; and (4) the right to privacy in photographs and films.[132] Moral rights are rather weak in the UK as they can only be enforced if they are asserted in writing. However, attribution is a common condition of many data licence agreements. Researchers are taught to reference all materials used regardless of moral

---

[127] Stef Van Gompel, (pp. 670-671).

[128] Alexander Ross, 'Copyright works: seeking the lost'. *Entertainment Law Review*, 25 (3) (2014), 104-107; Eleonora Rosati, 'The orphan works provisions of the ERR Act: are they compatible with UK and EU laws?' *European Intellectual Property Review*, 35 (12) (2013), 724-740.

[129] Graham M. Dutfield and Uma Suthersanen, 'The innovation dilemma: intellectual property and the historical legacy of cumulative creativity', *Intellectual Property Quarterly*, 4 (2004), 379-421.

[130] Matteo Migheli and Giovanni B. Ramello, 'Open access, social norms and publication choice', *European Journal of Law & Economics*, 35 (2) (2013), 149-167.

[131] Migheli and Ramello, (p. 149-150).

[132] The Copyright, Designs and Patents Act 1988, ss.77- 89 <http://www.legislation.gov.uk/ukpga/1988/48/part/I/chapter/IV> [accessed 9 August 2015];

during the 1928 revision in Rome attended by thirty-six signatories, Article $6^{bis}$ – 'Moral Rights' was added to the *Berne Convention* which confers minimum international standards for all signatories to abide by: 'Berne Convention: For the Protection of Literary and Artistic Works', *World Intellection Property Organization (WIPO) Website* <http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html> [accessed 9 August 2015].

rights. Researchers therefore may often consider attribution more in terms of social norms than in terms of legal responsibility.

Key stakeholders' awareness of pertinent legal issues is crucial for the effective re-usage of academic research data in a digital age. Higher education institutions must keep pace with the continual changes in legal practice; the introduction of the freedom of information exception for research and copyright exemption for data/text mining for researchers are two such recent examples.[133] Moreover, while rights management is also essential, researchers often do not fully comprehend the legalities.[134]

Legal services are provided by higher education institutions to support staff through legal rights clearance and management. Other higher education departments, such as libraries, repositories and innovation services can also provide relevant guidance through consultation, training sessions and web pages.[135] However, many of these departments often do not have the capacity to advise staff on every single dataset.[136] Furthermore, it is unclear how often advice and consultation is requested.[137]

Legal support for academic research also comes from external organisations, such as open licensing providers and the UK Intellectual Property Office (IPO). Since 2005, the IPO offers five model research collaboration agreements as part of the Lambert Toolkit on its website, to facilitate collaboration between businesses and universities conducting academic research.[138] However, according to an independent study conducted in 2013, this toolkit appears underused: 'less than half (45%) of the overall survey sample have used any part of the toolkit'.[139]

---

[133] Dave Hughes, 'The new research exemption - where information law and IP collide', *Freedom of Information*, 10 (6) (2014), 5-7; Intellectual Property Act 2014, section 20 'Freedom of information: exemption for research' <http://www.legislation.gov.uk/ukpga/2014/18/section/20/enacted> [accessed 9 August 2015].

[134] Jane Secker and Maria Bell, 'Copyright? Why would I need to worry about that? The challenges of providing copyright support for staff', *Legal Information Management*, 10 (3) (2010), 166-170 (, p. 166).

[135] Secker and Bell, (p. 166).

[136] The following article focuses on the challenges faced by libraries providing guidance on copyright: Dunstan Speight and Jennifer Darroch, 'Clarifying copyright', *Legal Information Management*, 12 (3) (2012), 209-213.

[137] Secker and Bell.

[138] 'Guidance: Lambert Toolkit', *UK Intellectual Property Office: UK Government Website* <https://www.gov.uk/lambert-toolkit> [accessed 9 August 2015].

[139] Elaine Eggington, Rupert Osborn and Claude Kaplan, 'Collaborative Research between Businesses and Universities: The Lambert Toolkit 8 Years On', An independent report commissioned by the Intellectual Property Office (IPO) in collaboration with AURIL, CBI, PraxisUnico & TSB and carried out

In consequence, despite greater access to and choice of open legal tools (e.g. open licences), people may fail to utilise such methods correctly due to limited legal awareness across the academic research community. This thesis therefore does not focus on the legal minutiae such as legal repeals and amendments (for instance many commentators have called for copyright reform).[140] It instead examines these pertinent legal issues on a practical and consciousness-raising level, for instance: (1) are data licences used within the case studies – if so what type; (2) what other legal issues arise within a specific data context (e.g. data protection issues where sensitive and personal data are being collected); and, (3) the significant gaps in legal awareness and support.

### 2.3.4 Traceable data: provenance

Thus far, data accessibility, trust and authorisation have been shown as vital for effective re-usage. However, the rich legal, technological and socio-cultural protocols and practices that pertain to the collection, management and re-usage of a particular dataset need to be recorded and therefore traceable. Without such assurances available to consult, the re-usage of an academic research dataset is significantly diminished.

The "etymology [for provenance] is the French verb 'provenir', which means to come forth, originate."[141] The term provenance is defined by the OED as:

> […] the place of origin or earliest known history of something […] a record of ownership of a work of art or an antique, used as a guide to authenticity or quality […] Origin: late 18th century: from French, from the verb *provenir* 'come or stem from', from Latin *provenire*, from *pro-* 'forth' + *venire* 'come'[142] [OED's italics]

---

by IP Pragmatics Limited (2013) *UK Intellectual Property Office: UK Government Website* (p. 20) <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/311757/ipresearch-lambert.pdf > [accessed 9 August 2015].

[140] Sun; Silver and Anderson; Christina J. Angelopoulos, 'Modern intellectual property legislation: warm for reform', *Entertainment Law Review*, 19 (2) (2008), 35-40; David Gee, 'Should librarians and information professionals be content with current UK copyright law?' *Legal Information Management*, 8 (3) (2008), 204-213.

[141] Luc Moreau, 'The Foundations for Provenance on the Web', *Foundations and Trends in Web Science*, 2 (2-3) (2010), 99-241, (p. 18) <http://dx.doi.org/10.1561/1800000010> (ePrint version <http://eprints.soton.ac.uk/271691/> [accessed 9 August 2015]).

[142] 'Definition of provenance', *Oxford English Dictionaries Website* <http://www.oxforddictionaries.com/definition/english/provenance?q=provenance> [accessed 9 August 2015].

From this dictionary definition, it is clear that the concept of provenance has a long-history, and has been a valuable tool that has mainly been used by art curators, archivists, archaeologists and rare book librarians for hundreds of years.[143] Provenance carries different meanings for stakeholders in these disciplines, as James Cheney and others state:[144]

> The concept of "provenance" originates from the art and archiving worlds, where it refers to information about the creation, chain of custody, modifications or influences pertaining to an artifact. [/] […][/][…][/]A more concrete answer to the question "what is provenance for *digital* artifacts" is to look at features or applications of current computer systems that appear related to history tracking, logging, integrity, authenticity, or error recovery.[145] (Cheney and others italicised emphasis.)

Provenance information is therefore vital to establish an item's authenticity and integrity. For example, provenance is used to aid the verification of whether a painting was truly created by Cézanne or is a forgery. It is also useful for unnamed sources, such as manuscripts and encyclopaedias.

While the concept of provenance remained essential in a print age, Alexandra Gillespie describes the medieval research users' lack of need for recorded provenance information:

> In a world in which twenty volumes was a 'riche' collection, the materials in which books were clad were liable to be rather empty and uninformative. It would take the Clerk some time to save enough money to buy and bind twenty tomes, but when he did he would probably not need a spine label, catalogue, index, or title page to remember some essential points about each (author and/or content, provenance, cost, and value to the reader for instance).[146]

Citation is not enough to convey the level of information needed to determine the quality and ultimately the re-usability of a dataset. According to Kiran-Kumar Muniswamy-Reddy and others, a dataset has:

---

[143] James Cheney and others, 'Provenance: A Future History', *The 24th International Conference on Object Oriented Programming, Systems, Languages and Applications (OOPSLA)*, Date: 25-29 October 2009, Location: Florida, USA, pp. 1-8, (p. 4) <http://people.seas.harvard.edu/~chong/pubs/onward09-provenance.pdf> [accessed 9 August 2015]; Shelley Sweeney, 'The Ambiguous Origins of the Archival Principle of "Provenance"', *Libraries & the Cultural Record* , 43 (2) (2008), 193-213 (p. 193-5) <http://dx.doi.org/10.1353/lac.0.0017>

[144] Sweeney, p. 193.

[145] Cheney and others, p. 4.

[146] Alexandra Gillespie, *Print Culture and the Medieval Author: Chaucer, Lydgate, and their Books 1473-1557* (Oxford: Oxford University Press, 2006), p. 2.

> two critical components: what it is (its contents) and where it came from (its ancestry). Traditional work in storage and file systems addresses the former: storing information and making it available to users. Provenance addresses the latter.[147]

Prior to the Web, but rooted firmly in the digital age, David A. Bearman and Richard H. Lytle published an article in 1985-6 that recognised the value of increased importance provenance for electronic data:

> The task of managing information in organisations is becoming more challenging as the organizations become larger and more complex, and as information technologies and general societal development increase the volume and sophistication of available information.[148]

In the digital age, the concept of provenance is increasingly utilised by the scientific community to provide information about academic research data; due to the increase in 'data-driven scientific investigations' and the vast amount of data which resides in databases.[149] With the amount of academic research data released on the Web growing rapidly, and the greater assimilation of data from different sources of mixed provenance, it is becoming increasingly challenging for research users to decide on which datasets are most reliable.[150] Tope Omitola and others raise the importance of automatically highlighting the origins and restrictions of a dataset to a prospective research user:

> On the Web, a user may be confronted with a potentially large number of diverse data sources of variable maturity or quality, and selecting the high quality data that are pertinent for their uses may be difficult. They would like to have mechanisms to automatically determine whether a web document or resource can be used, based on the original source of the content, the licensing

---

[147] Kiran-Kumar Muniswamy-Reddy, Peter Macko and Margo Seltzer, 'Provenance for the Cloud', *FAST'10 Proceedings of the 8th USENIX conference on File and storage technologies* (2010) (p. 1) <https://www.usenix.org/legacy/events/fast10/tech/full_papers/muniswamy-reddy.pdf?CFID=450573987&CFTOKEN=30484129> [accessed 9 August 2015].

[148] David A. Bearman and Richard H. Lytle, 'The Power of the Principle of Provenance', *Archiveria*, 21 (Winter 1985-1986), 14-27 (p. 14) <http://journals.sfu.ca/archivar/index.php/archivaria/article/view/11231/12170> [accessed 4 July 2015].

[149] Yogesh L. Simmhan and others, 'Performance Evaluation of the Karma Provenance Framework for Scientific Workflows' in *Provenance and Annotation of Data: International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006, Revised Selected Papers*, ed. by Luc Moreau and Ian Foster (Springer Berlin Heidelberg, 2006) 222-236 (p. 222) <http://dx.doi.org/10.1007/11890850_23>

[150] Clifford A. Lynch, 'When Documents Deceive: Trust and Provenance as New Factors for Information Retrieval in a Tangled Web', *Journal of the American Society for Information Science and Technology*, 52 (1) (2001), 12-17 <http://dx.doi.org/10.1002/1532-2890(2000)52:1<12::AID-ASI1062>3.0.CO;2-V>

> information associated with the resource, and any usage restrictions on that content, etc.[151]

In addition to this, Luc Moreau and Paul Groth state provenance is not just used by art curators anymore, but has become a fundamental part of the digital age:

> The World Wide Web is now deeply intertwined with our lives, and has become a catalyst for a data deluge, making vast amounts of data available online, at a click of a button. With Web 2.0, users are no longer passive consumers, but active publishers and curators of data. […] Provenance is no longer seen as a curiosity in art circles, but it is regarded as pragmatically, ethically, and methodologically crucial for our day-to-day data manipulation and curation activities on the Web.[152]

Due to the long-established use of provenance primarily within the arts, humanities and archival communities, and its reawakening within computer science and other scientific disciplines, there is no single comprehensive definition for provenance. For instance, Jie Yuan and others describe data provenance as recording the sources and a set of processing steps applied to sources.[153] Tope Omitola and others state that provenance: 'describes how an object came to be in its present state, and thus, it describes the evolution of an object over time'.[154] Antonio Badia contends that provenance metadata relates to: 'the origin of the data: where it comes from, how and when it was obtained, and any relevant conditions that might help determine how it came to be in its current form'.[155] Yogesh L. Simmhan and others maintain that provenance metadata: 'tracks the steps by which the data was derived […]'.[156]

Sudha Ram and Jun Liu's generic W7 Model proposes seven key interrelated elements that are required for a robust record of data provenance, which encompasses:

---

[151] Tope Omitola and others, 'Tracing the Provenance of Linked Data using voiD', *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, Date: 25-27 May 2011, Location: Sogndal, Norway (p.1) <http://doi.acm.org/10.1145/1988688.1988709>

[152] Luc Moreau and Paul Groth, *Provenance: An Introduction to PROV* (Morgan & Claypool Publishers, 2013) p. vi. eBook <http://dx.doi.org/10.2200/S00528ED1V01Y201308WBE007>

[153] Jie Yuan, Jianya Gong and Mingda Zhang, 'A Linked Data Approach for Geospatial Data Provenance', *IEEE Transactions on Geoscience and Remote Sensing*, 51 (11) (2013), 5105-5112 (p. 5105) <http://dx.doi.org/ 10.1109/TGRS.2013.2249523>

[154] Omitola and others, p. 2.

[155] Antonio Badia, 'Evaluating Source Trustability with Data Provenance: A Research Note', *Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI)*, Date: 11-14 June 2012, Location: Washington D.C., USA, pp. 129-131 (p.129) <http://dx.doi.org/10.1109/ISI.2012.6284145>

[156] Yogesh L. Simmhan, Beth Plale and Dennis Gannon, 'A Survey of Data Provenance in e-Science', *SIGMOD Record*, 34 (3) (2005), 31-36 (p. 31) <http://www.sigmod.org/publications/sigmod-record/0509/p31-special-sw-section-5.pdf> [accessed 9 August 2015].

(1) what – the 'sequence of events […] that affect a data object during its lifetime'; (2) when –'a set of timestamps […] associated with various provenance events'; (3) where – 'a set of locations […] where various events happen'; (4) how – 'actions that lead to the occurrence of an event' such as preconditions, methods, inputs, outputs and sources; (5) who – 'people and/or organisations that cause events'; (6) which – '*devices* used in data creation, analysis and transformation'; and, (7) why – 'a set of reasons […] for various provenance events'.[157] The W7 model is a useful account of some of the key provenance parameters; however, it is limited as it does not expressly address legal rights ownership and clearance.[158]

There are also a number of well-known metadata systems to address traceability issues, such as: Dublin Core – an open organisation which supports metadata design and best practice; and, DCAT – an RDF vocabulary used to describe datasets held in data catalogues to facilitate interoperability.[159] More recently Luc Moreau and others worked on the PROV-DM model – a conceptual data model including provenance constraints and notation – which became a W3C recommendation on 30 April 2013.[160] There is no doubt that provenance is a vital part of data creation, storage, maintenance and retrieval. Due to the different definitions of provenance this thesis utilises its own definition (refer to glossary):

> A digitally accessible record pertaining to a specific academic research dataset, including information about its origins, development, versions, legal, technological and socio-cultural frameworks.

In summary, the growing number of researchers and datasets has positioned provenance at the core of effective academic research data re-usage in a digital age. Chapters 4-6 will further explore questions of provenance.

---

[157] Sudha Ram and Jun Liu, 'Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling' in *Active Conceptual Modeling of Learning: Next Generation Learning-Base System Development*, ed. by Peter P. Chen and Leah Y.Wong (Springer Berlin Heidelberg, 2007) pp. 17-29. Springer eBook <http://dx.doi.org/10.1007/978-3-540-77503-4_3>

[158] John Mark Ockerbloom, 'Copyright and Provenance: Some Practical Problems', *IEEE Computer Society Technical Committee on Data Engineering*, 30 (4) (2007), 51-58 <http://repository.upenn.edu/cgi/viewcontent.cgi?article=1051&context=library_papers> [accessed 9 August 2015].

[159] 'Data Catalog Vocabulary' (DCAT), *W3C Website* <http://www.w3.org/TR/vocab-dcat/> [accessed 9 August 2015]; *Dublin Core Website* <http://dublincore.org/> [accessed 9 August 2015].

[160] 'PROV-DM: The PROV Data Model', *W3C Website* <http://www.w3.org/TR/prov-dm/> [accessed 9 August 2015].

### 2.3.5 Human-readable, machine-readable and machine-understandable data

A Web of linked documents, data, information and knowledge has been made digitally accessible on an unprecedented global scale. Evermore complex packets of academic research data are now possible (e.g. through data mining techniques), compared with the print medium. However, the Web is still in its infancy. While large amounts of academic research data are digitally accessible, a significant portion of these data are not yet machine-understandable.

There is a distinction to be drawn between machine-readability and machine-understandablity.[161] Machine-readability is where a machine can parse the data and provide an unambiguous structure (e.g. HTML and XML are machine-readable), but cannot provide any inherent meaning in the structure (i.e. tag *y* means *x*). Machine-understandability is where a machine not only provides an unambiguous structure, but also its inherent meaning (e.g. RDF is machine-understandable). There is potential for more value to be derived from automatically joining together, and being able to query, larger numbers of related interoperable and integrated datasets that are stored in distributed systems.

Machine-understandablity is variably known as the Semantic Web and Web of Linked Data. In general, the Semantic Web community uses expressive languages that permit inference for reasoning, such as OWL and RDF Schema. In contrast, the Web of Linked Data community sometimes utilises no inference, but aims to join up certain parts of separate datasets using hyperlinks. However, the Web of linked documents has yet to fully embrace machine-understandability.

Since the First World Wide Web Conference was held in 1994, Tim Berners-Lee has envisioned a Web of linked data.[162] Moreover, in 2001, the concept of the Semantic Web was first published by Tim Berners-Lee, James Hendler and Ora Lassila in

---

[161] For further information on machine-readability refer to: James Hendler, 'Developers: A Primer on Machine Readability for Online Documents and Data', 24 September 2012, *USA Government Data Website: Developers* <https://www.data.gov/developers/blog/primer-machine-readability-online-documents-and-data> [accessed 9 August 2015].

[162] *The First World Wide Web Conference*, Location: CERN, Geneva, Switzerland (25-27 May 1994) <http://www94.web.cern.ch/WWW94/Welcome.html> [accessed 9 August 2015]; Tim Berners-Lee, Wendy Hall and Nigel Shadbolt, 'The Semantic Web Revisited', *IEEE Intelligent Systems*, May/June (2006), 96-101 (p. 96) <http://eprints.ecs.soton.ac.uk/12614/1/Semantic_Web_Revisted.pdf> [accessed 9 August 2015]. ePrint.

*Scientific American*.[163] Tim Berners-Lee and others describe the difference between the Web of linked documents and the Web of linked data:

> The glue that holds together the traditional document Web is the hypertext links between HTML pages. The glue of the data Web is RDF links. An RDF link simply states that one piece of data has some kind of relationship to another piece of data.[164]

Alexander León and others further explain the function of linked data:

> Technically, Linked Data is about employing the RDF language and the HTTP protocol to publish structured data on the Web and to effectively connect data between different data sources through deferenceable URIs […] Data published this way is machine-readable and its semantics is explicitly defined.[165]

A number of diverse organisations are already employing linked data, including the British Broadcasting Corporation (BBC), Ordnance Survey, the UK government and the University of Southampton.[166] The Linked Data Cloud is maintained by Richard Cyganiak and Anja Jentzsch and tracks the number of datasets published in a linked data format.[167] The first linked data cloud was published on 1 May 2007; only twelve linked datasets were recorded.[168] The most current survey, published on 30 August 2014, captured 570 linked data records.[169]

While there is increased focus on the generation of machine-understandable data, the Web of linked data is yet to be fully achieved. As with the early stages of the

---

[163] Tim Berners-Lee, James Hendler and Ora Lassila, 'The Semantic Web', *Scientific American*, May (2001), 34-43 < http://www.scientificamerican.com/article/the-semantic-web/> [accessed 9 August 2015].

[164] Christian Bizer and others, 'Linked Data on the Web', *Proceeding of the 17th international conference on World Wide Web*, Session: Workshops, Date: 21-25 April 2008, Location: Beijing, China, 1265-1266 (p. 1265) <http://doi.acm.org/10.1145/1367497.1367760>

[165] Alexander León and others, 'Geographical Linked Data: a Spanish Use Case', *Proceedings of the 6th International Conference on Semantic Systems, Session: Pragmatic Web, Conference track: triplification challenge*, Date: 1–3 September 2010, Location: Messe Congress Graz, Austria, 36, 1-3 (p. 1) <http://doi.acm.org/10.1145/1839707.1839753>

[166] Oliver Bartlett, 'Linked Data: Connecting together the BBC's Online Content', *BBC Blog*, 19 February 2013 <http://www.bbc.co.uk/blogs/internet/posts/Linked-Data-Connecting-together-the-BBCs-Online-Content> [accessed 9 August 2015]; 'OS OpenData', *Ordnance Survey Website* <http://www.ordnancesurvey.co.uk/business-and-government/products/opendata-products.html> [accessed 9 August 2015]; 'University of Southampton Open Data Service', *University of Southampton Website* <http://data.southampton.ac.uk/> [accessed 9 August 2015].

[167] Richard Cyganiak and Anja Jentzsch, *Linked Data Cloud Website* <http://lod-cloud.net/> [accessed 9 August 2015].

[168] *Ibid.*

[169] *Ibid.*

Web, the number of users increased when tools were released that were easy for people with non-technical skills to use, as Deborah L. McGuinness states: [170]

> In the early days of the web, HTML pages were generated by hand. The pages contained information about how to present information on a page. Early adopters took to the web quickly since it provided a convenient method for information sharing. Arguably, the generation of tools for machine generation and management of web pages allowed the web to really take off. Tool platforms allowed non-technical people to generate and publish web pages quickly and easily. The resulting pages typically included content and display information and targeted human readers (rather than targeting programs or automatic readers).[171]

While this thesis is unable to offer definitive answers on the future development of the Web, it appears likely that academic research data will become more machine-understandable and therefore increasingly interoperable, integrated and queryable. This would have significant impacts for data accessibility, trust, authorisation and provenance, and would ultimately further re-shape knowledge transfer. Machine-understandable data is therefore not a new concept on the Web.

This thesis is only able to provide an extremely brief and high level overview of the Web of linked data. However, just as ahistoricism is problematic when considering knowledge transfer, it is also unwise to overlook ways in which digital culture might mature. In order for academic research data re-usage to remain sustainable, there must be cognisance of and adaption to future change. It is expected that greater automation and seamless work flows of machine-understandable data might pose new challenges for provenance, legal, technical and socio-cultural frameworks.

Chapter 4 and Chapter 6 provide an insight into working models of machine-readable provenance metadata and academic research data. Although at the time of the interviews in 2012, MEDIN were aware about the benefits of linked data there was limited funding for its implementation. Whereas, in Chapter 5, the Southampton Chemical Crystallography Group (SCCG) are actively producing machine-understandable provenance metadata within its archive and laboratory notebook work.

While this thesis acknowledges that in practice (Chapters 4 and 6) machine-understandable provenance metadata and academic research data are not being fully

---

[170] Deborah L. McGuinness, 'Ontologies Come of Age' in *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*, ed. by Dieter Fensel and others (Massachusetts: MIT Press, 2003) <http://www-ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-mit-press-(with-citation).htm> [accessed 9 August 2015].

[171] *Ibid.*

implemented, it does not underestimate the potential for and growing application of machine-understandable technologies in the future. In consequence, understanding how the potential for greater machine-understandable data may impact on future re-usage of academic research data is essential to provide resilient principles of best practice (Chapter 7) and key areas for future work (Chapter 8). Better comprehension of machine-understandable data and its possible role for successful academic research data re-usage in a digital age is therefore a vital consideration for the thesis conclusions, which aims not only to model best practice now but for a maturing Web.

## 2.4 Literature review: summary

This chapter began by reviewing a motivating example of worst practice – the case of Hwang and others – to map how the processes which should have prevented its occurrence, failed. This example revealed three key areas for concern whilst modelling best practice: (1) safeguarding diverse academic research from a multitude of sources and for a variety of end users; (2) overcoming issues of multi-authorship in highly collaborative research environments; and, (3) facilitating the re-usage of academic research data gathered from human participants which are amongst the most difficult to release or fully disclose as open data; especially where those data are personal and, in some cases, extremely sensitive.

The literature review then evaluated the historical and contextual development of the best practice safeguards – data accessibility, trust, authorisation and traceability – that together amount to good quality academic research data re-usage. This required examination of literature from a multitude of disciplines, which have raised an extensive range of issues concerning the robustness, reliability and re-usability of academic research data.

However, there is currently no foremost authority that wholly encapsulates: (a) how the re-usage of academic research data has been re-versioned for a digital age; and, (b) what best practice principles are required to guarantee its robustness across the sciences, social sciences and humanities both now and in the future. The key literature is often disjointed and scattered within a number of disciplinary domains that require a familiarisation with specialist terminology, therefore making it more difficult to determine which of the most influential texts should take principal position within the literature review. There are a number of stand-out texts however.

From the web and internet science literature, Luc Moreau's work shows the importance of provenance not only within the digital age, but as part of Web Science research, and Sudha Ram and Jun Liu's generic W7 Model offers a useful introduction to the scope of provenance. From the legal literature, Kansa and others' valuable article recognises the diverse nature of academic research data – a concept that is sometimes lacking within legal scholarship – and that this diversity may attract a number of diverse rights. From the humanities literature, the works of both Elizabeth L. Eisenstein and Adrian Johns explore the impact of the printing press on knowledge transfer beyond technological factors. Although not strictly within the humanities literature, Katherine Gross and Gary Mittleback raised a number of interesting ethical questions about the integrity of science.

This chapter has sought to challenge ahistorical assumptions about knowledge transfer, where some research neglects to recognise that many of the challenges facing it in a digital age are similar to many issues encountered in the print age. For instance, it is clear that the open movements of the digital age bear a striking resemblance to the republic of letters of the print age.

Overall the literature review has revealed a wealth of (provenance) metadata standards, ethical codes of practice and guidance, legal rights, licensing, and technologies. These research protocols have a combined impact on modelling best practice for re-usage of academic research data in a digital age. However, the aim of this thesis is not to add to these important but burgeoning numbers of research protocols by offering a new moral code, a new provenance metadata standard, a new technical system for knowledge transfer or amendments to legislation. Rather this thesis focuses on issues arising from existing practice, therefore recognising that there is not only an intellectual need but a pragmatic requirement for research in this area. Moreover, as a significant amount of this research focuses on scientific data, this thesis also aims to raise the profile of quality academic research data re-usage within the social sciences and humanities.

The two primary and three secondary research questions addressed by this thesis have already been identified within Chapter 1, section 1.3.[172] A key motivation for this

[172] The two primary research questions are as follows: (1) what makes for excellent academic research in a digital age? (2) How should best practice for academic research data re-usage be modelled both now and in the future? What can be learnt from longstanding practices? The three secondary research questions are as follows: (3) How should diverse types of academic research data from multiple

thesis is both to examine: the extent in which the re-usage of academic research data has been re-versioned for a digital age; and, the best practice principles required for its successful re-usage across the sciences, social sciences and humanities both now and in the future. This interdisciplinary literature has shown that an intellectual study of academic research data re-usage practices is not enough. In order to produce useful best practice principles pertinent to the pragmatic requirements of the academic research community across the sciences, social sciences and humanities, this thesis focuses on the practicalities of academic research data re-usage through three cases studies in Chapters 4-6.

### 2.4.1 Mapping from Chapter 2 to the case studies in Chapters 4-6

This chapter has already began to identify a number of significant narrative threads, which link the academic research data re-usage issues raised by the interdisciplinary literature review with the case studies in Chapters 4-6. It has evaluated the key print and therefore e-print processes for good quality data re-usage within four key topics that emerge from the literature: (1) data accessibility, (2) trust, (3) authorisation, and (4) traceability. Each of these topics is composed of a variety of interdisciplinary issues. For instance, data accessibility is not only predicated on technical search and storage capabilities, but individuals' attitudes to sharing data (the wider socio-cultural context). Another example is provenance which needs to capture wider legal, socio-cultural, and technological information e.g. attribution, re-usage permissions, ethics clearance and formats. Given the interdisciplinary nature of these topics, the focus of the case studies will be on their four fundamental areas of support: (i) provenance frameworks and issues; (ii) legal frameworks and issues; (iii) technological frameworks and issues; and, (iv) socio-cultural frameworks and issues.

This thesis therefore takes forward the key safeguards raised in this chapter through three data generation arenas normally not considered together: chemistry, marine environmental sciences, and modern languages. The three case studies focus on five working data platforms within the marine environmental sciences, chemistry and

---

originators, contexts and sources be safeguarded for a wide set of research users? (4) How should academic research data generated as part of collaborative research be treated in order to balance a range of difficult communal and continuity problems? (5) How should maintenance of and access to sensitive and personalised data be treated in order to balance a range of difficult problems with permissions, data protection and confidentiality?

modern languages: MEDIN, eCrystals, LabTrove, FLLOC and SPLLOC. A data platform is defined as (refer to glossary):

> An online collection of re-usable academic research data which operates independent to academic publications (such as academic journals), including laboratory/project-based data repositories, institutional and (inter)national data archives.

Each case study has been selected to confront one of the three main areas of concern raised by the case of Hwang and others, to ensure potential worst practice is mitigated. Chapter 4 focuses on how diverse types of academic research data from multiple originators, contexts and sources can be safeguarded for a wide set of research users. Chapter 5 explores how the maintenance of and access to academic research data generated by multiple originators can balance a range of difficult problems with mutual permissions and joint authorship. Chapter 6 addresses how maintenance of and access to sensitive and personalised data can balance an assortment of challenging issues with permissions, data protection and confidentiality.

In summary, the provenance, legal, technological and socio-cultural frameworks supporting these three cases studies (and thus facilitating the creation, maintenance and re-usage of accessible, trusted, authorised and traceable academic research data) are explored through consultation of their websites, secondary literature, and eighteen interviews with individuals connected to these models. These participants have a range of expertise in academia, data policy, data management, technology and law. This case study and semi-structured interview approach is now fully explained in Chapter 3.

# Chapter 3: Methodology

## 3.1 Methodology: Introduction

For some disciplines, such as law and within the humanities, a methodology chapter is not a separate requirement. The methodology is often integrated within the wider synthesis. However, this distinct methodology chapter is important for three principal reasons. Firstly, it locates this thesis within the interdisciplinary field of web science in order to show the reasons why it is well-placed to confront the range of complex issues which impact on academic research data re-usage. Secondly, this chapter outlines the rationale, justifications for and potential limitations of the chosen methodological approach – an interdisciplinary literature review, case studies and semi-structured interviews. Finally, as this chapter is a case in point for good academic research data gathering practice, it provides a pre-amble to Chapter 6 – the case study that focuses on the re-usage of personal and sensitive academic research data.

While the semi-structured interview data collected are not highly personal or sensitive, the participants are still discussing aspects of their employment and expressing personal opinions. It is therefore vital that the ethical processes and clearance undertaken by this thesis are reviewed. The chapter examines the processes involved with the lifecycle (collection, storage, use and deletion) of the academic data generated by this thesis through eighteen semi-structured interviews. The crucial aspect of this chapter lies in its appraisal of how this thesis manages its own data accessibility, trust, authorisation, and traceability (concepts emphasised by Chapter 2). What provenance metadata, legal, ethical, technological and socio-cultural frameworks are being employed and need to be made transparent or accessible?

## 3.2 Web Science

Web Science is a relatively new emerging field of interdisciplinary study (this thesis forms part of the first cohort of Web Science research projects at the University of Southampton).[173] It aims not only to understand the technological architecture and

[173] For more background information on Web Science refer to the following: James Hendler and others, 'Web science: an interdisciplinary approach to understanding the web', *Communications of the ACM*, 51

impacts of the Web, but its wider societal and cultural effects. Chapter 2 shows that pre-digital and digital technologies are influenced by their cultures and societies. For instance, Elizabeth L. Eisenstein and Adrian Johns' separate works could be labelled as part of printing press science. Moreover, areas such as digital humanities, information science and IT law (although largely separate) are already at the forefront of addressing issues manifesting within the digital age. Despite this, Chapter 2 also demonstrates that this wealth of material is often theoretical, trapped within disciplinary silos, too policy-driven, and/or lacking historical context.[174]

In consequence, the field of Web Science does not simply join up existing research areas. It is confronting a number of Web related issues from a variety of disciplinary angles by critical thinking across domains. Through this interdisciplinary comprehension, this thesis is able to provide a new synthesis of existing research by joining up and filling the gaps between previously unconnected areas in ways hitherto not undertaken.[175]

In order to understand the key (inter)disciplinary issues arising from modelling best practice for academic research data re-usage in a digital age, this thesis draws on expertise from three domains: the humanities, law, and web and internet science. A humanities approach unpacks longstanding historical and contextual development of data re-usage. A legal approach is required to review the current challenges with

---

(7) (2008), 60-69 <http://dx.doi.org/10.1145/1364782.1364798>; 'Web Science', *University of Southampton Website* <http://www.southampton.ac.uk/webscience> [accessed 9 August 2015]; *Web Science Trust Website* <http://webscience.org/> [accessed 9 August 2015]; 'Web Science: how the Web is changing the world', *Future Learn Website* <https://www.futurelearn.com/courses/web-science> [accessed 9 August 2015].

[174] The thesis author presented an extended abstract on interdisciplinary research: [unpublished] Laura German, 'Overcoming Methodological Silos through Interdisciplinary Research: The Case of the Web Scientist' at *London Centre for Social Studies PhD Conference 2013: Methodological Choices and Challenges*, King's College London, 19 April 2013 <http://socialstudies.org.uk/Events/Conferences/6266/1st-LCSS-PhD-Conference-2013-Methodological-Choice> [accessed 9 August 2015].

[175] Refer to the following literature for further information on the historical development and contextual development of interdisciplinarity: Julie Thompson Klein, *Humanities, Culture, and Interdisciplinarity: The Changing American Academy* (New York, NY: State University of New York Press, Albany, 2005) Google ebook; Elizabeth Dzeng, 'How to inspire interdisciplinarity: lessons from the collegiate system', *Guardian*, 15 March 2013 <http://www.theguardian.com/higher-education-network/blog/2013/mar/15/interdisciplinary-academic-universities-research> [accessed 9 August 2015]; 'More about interdisciplinarity', *University College London Website* <http://www.ucl.ac.uk/basc/faq/interdisciplinarity> [accessed 9 August 2015].

existing legal practice in the context of academic data re-usage within the UK higher education sector. A web and internet science approach provides technological insight into a maturing Web, and its potential implications for facilitating and modelling best practice.

As this thesis is co-supervised by three faculties ((1) Humanities, (2) Business, Law and Art, and (3) Physical Sciences and Engineering) and written for an interdisciplinary audience, a glossary is provided to avoid any ambiguity where certain terms are discipline specific or have distinct meaning within different disciplines. It is now important to outline the methodological approach undertaken by this thesis for research transparency and to acquaint those unfamiliar to the use of case studies and semi-structured interviews.

## 3.3 Case studies

There is an intellectual and pragmatic need to address two key issues: (1) what makes for excellent academic practices; and, (2) what principles should inform how best practice is modelled, now and in the future. Although self-evident, academic research data re-usage is a practical activity carried out by individuals in the course of their research across all disciplines. This thesis therefore needs to engage with the realities of academic research data re-usage by examining working models across the sciences, social sciences and humanities.

Chapter 2 has already shown the merits of a case study approach by Hwang and others as an indicative and hence motivating example of worst academic practice in a digital age.[176] This initial case study helped to unlock three main areas for concern on

[176] For further background information on the potential scope, advantages and limitations of the case study approach refer to the following literature: Peter Halfpenny, 'The analysis of qualitative data', *The Sociological Review*, 27 (4) (1979), 799-827 (p. 799). <http://dx.doi.org/10.1111/j.1467-954X.1979.tb00361.x>; Bent Flyvbjerg, 'Five Misunderstandings About Case-Study Research', *Qualitative Inquiry*, 12 (2006), 219-245 <http://dx.doi.org/10.1177/1077800405284363>; D. Silverman, 'Qualitative research: meanings or practices?' *Information Systems Journal*, 8 (1) (1998), 3-20 <http://dx.doi.org/10.1046/j.1365-2575.1998.00002.x>; Dilanthi Amaratunga and David Baldry, 'Case study methodology as a means of theory building: performance measurement in facilities management organisations', *Work Study*, 50 (3) (2001), 95-105 <http://dx.doi.org/10.1108/00438020110389227>; Catherine Pope, Sue Ziebland and Nicholas Mays, 'Analysing qualitative data', *British Medical Journal*, 320 (7227) (2000), 114-116 <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1117368/> [accessed 9 August 2015]; Robert K. Yin, *Case Study Research: Design and Methods*, 4th edn (London: SAGE

modelling best practice (these are re-iterated in sections 3.3.1-3.3.3), which are confronted by Chapters 4-6. A case study approach enriches a literature review by encompassing secondary literature (that is often published outside traditional academic publication channels e.g. codes of conduct and standards documentation). It also enables direct interaction with functioning data platforms through their websites.

In order to raise consciousness of the complexities around managing and releasing diverse types of academic research data on the Web, regardless of discipline or research methodology, this thesis has chosen three case studies across the sciences, social sciences and humanities. Chapter 4 focuses on the Marine Environmental Data and Information Network (MEDIN). Chapter 5 examines eCrystals and LabTrove from the field of chemistry. Chapter 6 focuses on the French Learner Language Oral Corpora (FLLOC) and the Spanish Learner Language Oral Corpora (SPLLOC) based within modern languages.

To maximise awareness of these complexities, MEDIN, eCrystals, LabTrove, FLLOC and SPLLOC manage different types of data which are delivered through a diverse range of platforms. All operate within the UK higher education sector (although MEDIN has a wider remit as the UK government mechanism for marine environmental data re-usage). This diverse range of case studies presents an opportunity to examine a variety of provenance metadata, legal, technological and socio-cultural frameworks and issues.

Each case study selected has direct connections to each of the three thesis supervisors (who are situated within three different faculties). Professor Stephen Saxby (Law) has links to MEDIN through the GeoData Institute at the University of Southampton, and co-authored a MEDIN report on data policy published in December 2010.[177] Professor Leslie Carr (Web and Internet Science) has connections to eCrystals through the EPrints software, and the series of collaborations between chemistry and

---

Publications, 2009), pp. 14-15; Kathleen M. Eisenhardt, 'Building Theories from Case Study Research', *Academy of Management Review*, 14 (4) (1989), 532-550 <http://www.jstor.org/stable/258557> [accessed 9 August 2015].

[177] Neil Pittam, Stephen Saxby and Chris Hill, 'Approaches to data policy in the marine sector', *'Marine Environmental Data and Information Network' (MEDIN) Final Project Report*, Version 1.1 (December 2010)
<http://www.oceannet.org/library/work_stream_documents/documents/medin_data_policy_study_rep_final_v1_1.pdf> [accessed 9 August 2015].

web and internet science at the University of Southampton (explained in Chapter 5). Finally, a number of Professor Mary Orr's (Modern Languages) departmental colleagues are/were involved with FLLOC and SPLLOC. A detailed overview of each of these case studies is provided by Chapters 4-6.

### 3.3.1 MEDIN case study

Chapter 4 focuses on the Marine Environmental Data and Information Network (MEDIN), which is the UK government mechanism for marine environmental data and information re-usage.[178] While MEDIN does not only manage academic research data, but marine environmental datasets collected by government bodies and private organisations, these aspects of the model fall outside the scope of this thesis. To re-iterate previous comments, as well as occurring across jurisdictions and disciplines, academic research data re-usage also happens across sectors and industries. Although this thesis focuses on the UK higher education sector, MEDIN is a reminder of the further complicated landscape in which knowledge transfer is situated. As MEDIN manages a large amount of diverse data types from multiple sources, it is well-placed to confront the following secondary research question: how should diverse types of academic research data from multiple originators, contexts and sources be safeguarded for a wide set of research users?

### 3.3.2 eCrystals and LabTrove case study

Chapter 5 centres on two existing data platforms: eCrystals and LabTrove. eCrystals is an online data archive for data generated through crystallography experiments.[179] LabTrove is an electronic laboratory notebook for recording scientific experiments and research data (refer to Chapter 5 for a more detailed overview of both these models).[180] Crystallography is a highly collaborative discipline and therefore this case study is suitably located to confront the following secondary research question: how should academic research data generated as part of collaborative research be treated in order to balance a range of difficult communal and continuity problems?

---

[178] *'The Marine Environmental Data and Information Network' (MEDIN Website)* <http://www.oceannet.org/> [accessed 9 August 2015].

[179] *eCrystals Website* <http://ecrystals.chem.soton.ac.uk/> [accessed 9 August 2015].

[180] *LabTrove Website* <http://www.labtrove.org/> [accessed 9 August 2015].

### 3.3.3 FLLOC and SPLLOC case study

Chapter 6 investigates FLLOC and SPLLOC, which are sister websites that offer access to second language acquisition research corpora.[181] A number of these datasets are collected from school children. A focus on life histories research further enriches this case study by providing access to **Dr K.** who has experience of collecting, managing, using and re-using data that are not just personal but highly sensitive. As a result this case study is well-positioned to consider the following secondary research question: how should maintenance of and access to sensitive and personalised data be treated in order to balance a range of difficult problems with permissions, data protection and confidentiality?

### 3.3.4 Existing case studies

Case studies provide a useful methodological mechanism by which the practical realities of knowledge transfer can be revealed and evaluated. A number of such studies have already been undertaken. For instance, on 9 April 2013, the Open Knowledge Foundation called for researchers across all disciplines to submit their personal experiences of working with open research data in the form of 200-500 words to be captured in a forthcoming Open Research Data Handbook.[182]

In June 2008, the Research Information Network commissioned the Share or not to Share Report, which comprised over a hundred interviews with researchers, data managers and data experts at UK higher education institutions to gather 'information on researchers' attitudes and data-related practices in six discrete research areas – astronomy, chemical crystallography, classics, climate science, genomics, and social and public health sciences – along with two interdisciplinary areas – systems biology and the UK's rural economy and land use programme'.[183] This case study was a robust

---

[181] *French Learner Language Oral Corpora (FLLOC) Website* <http://www.flloc.soton.ac.uk/> [accessed 9 August 2015]; *Spanish Learner Language Oral Corpora (SPLLOC) Website* <http://www.splloc.soton.ac.uk/> [accessed 9 August 2015].

[182] 'Open Research Data Handbook – Call for case Studies', 9 April 2013, *The Open Knowledge Foundation Blog Website* <http://blog.okfn.org/2013/04/09/open-research-data-handbook-call-for-case-studies/#sthash.2r4RbzuY.dpuf> [accessed 9 August 2015].

[183] Alma Swan and Sheridan Brown, 'To Share or not to Share: Publication and Quality Assurance of Research Data Outputs'. Report undertaken by Key Perspectives Ltd and commissioned by the 'Research Information Network' (RIN). RIN in association with the 'Joint Information Systems Committee' (JISC)

effort to gather specialist expertise from the people behind the process. However, by only focusing on researchers, data managers and data experts, it overlooked the other specialists that have a considerable role in academic research data re-usage, such as legal experts and IT professionals. While this study briefly highlighted some of the legal issues, it did not give a comprehensive overview of the socio-cultural, legal and technological issues. Moreover, this study focuses on the obstacles to sharing academic research data rather than modelling best practice.

In June 2012, the Royal Society used six case studies across the sciences to explore science as an open enterprise: the Astronomy and the Virtual Observatory; the Laser Interferometer Gravitational-wave Observatory project; the Scientific Visualisation Service for the International Space Innovation Centre; the UK Land Cover Map at the Centre for Ecology & Hydrology; the Global Ocean Models at the UK National Oceanography Centre; and, the Avon Longitudinal Study of Parents and Children.[184] While six case studies were used to demonstrate academic research data in practice, no interviews were used to gather specialist insight from the people behind the models and, as to be expected, these case studies focused only on the sciences and not the social sciences or humanities.

Case studies are employed by the Open Data Institute to emphasise the practical applications of open data; however these case studies focus on business models – start-up companies – built around the re-usage of open data. The Open Government Data blog also wants to use a number of case studies to show the practical re-usage of open government data across all sectors of society. Further, the DCC has case studies on its website for guidance.

---

and the 'Natural Environment Research Council' (NERC) (June 2008), *The Research Information Network Website* <http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs> [accessed 9 August 2015].

[184] 'Science as an open enterprise', The Royal Society Science Policy Centre Report, 02/12 (June 2012), *The Royal Society Website* <http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf> [accessed 9 August 2015]; 'Science as an open enterprise: case studies', *The Royal Society Website* <http://royalsociety.org/policy/projects/science-public-enterprise/case-studies/> [accessed 9 August 2015].

## 3.4 Semi-structured interviews

While case studies are effective for probing a wider set of secondary literature and exploring diverse data re-usage platforms, this approach is not enough. The answers to what makes for best academic practice and how it should be modelled ultimately lies with the people behind these models, as Anssi Peräkylä and Johanna Ruusuvuori state:

> By using interviews, the researcher can reach areas of reality that would otherwise remain inaccessible such as people's subjective experiences and attitudes. The interview is also a very convenient way of overcoming distances both in space and time; past events or faraway experiences can be studied by interviewing people who took part in them.[185]

Therefore, to enrich the interdisciplinary literature review and case studies, semi-structured interviews have been conducted with eighteen participants – six per case study (this is explained in section 3.4.1).

A semi-structured interview approach was chosen rather than structured or unstructured interview approaches.[186] A semi-structured interview approach requires that all interview participants are asked a set number of questions that are devised prior to the interviews (see section 3.4.5).[187] There is therefore a significant level of certainty that particular themes will be discussed through designed questions. This produces a partial, repeatable format which offers a level of consistency between interviews that is important for data comparison. This level of consistency is not provided by unstructured interviews, which have no particular repeatable format.

There is also a degree of flexibility which offers the potential to uncover relevant issues outside the scope of the current research themes and literature. There is time allocated to formulate further, relevant questions during the interviews which are tailored to the individual interview participant's role, experience, knowledge and

---

[185]Anssi Peräkylä and Johanna Ruusuvuori, 'Chapter 32: Analyzing Talk and Text' in *The SAGE Handbook of Qualitative Research*, 4th edn, ed. by Norman K. Denzin and Yvonna S. Lincoln (London: SAGE Publications, 2011), pp. 529-544 (p. 529). Google eBook.

[186]For further information about the strengths and weaknesses of structured and unstructured interviews refer to: Barbara DiCicco-Bloom and Benjamin F. Crabtree, 'The qualitative research interview', *Medical Education,* 40 (4) (2006), 314-321 (p. 314) < http://dx.doi.org/10.1111/j.1365-2929.2006.02418.x>; Tanya V. McCance, Hugh P. McKenna and Jennifer R.P. Boore, 'Exploring caring using narrative methodology: an analysis of the approach', *Journal of Advanced Nursing,* 33 (3) (2001), 350-356 (p. 351) <http://dx.doi.org/10.1046/j.1365-2648.2001.01671.x>

[187] Tom Wengraf, *Qualitative Research Interviewing* (London: SAGE Publications, 2001), p. 97.

interests.[188] Therefore, semi-structured interviews provide opportunities to ask further questions that are specific to the expertise of a particular interview participant, as C. Gibson states:

> In semi-structured interviewing the interviewer requires more focused information, and asks specific information to get it. The researcher opens the discussion, listens and prompts to guide the respondent.[189]

Structured interviews do not deviate from a set number of questions, and therefore do not capture this value-added information.

This thesis does not use surveys or questionnaires which, although useful methodological approaches, do not guarantee access to a known number of well-informed individuals with specific expertise. Furthermore, it would be difficult to tailor these to a participant's specific role. In contrast to semi-structured interviews, the interviewer is also unable to prompt the participant and offer to further define terms within questions where necessary.

### 3.4.1 Participant selection

To cover the wide range of provenance metadata, legal, technological and socio-cultural issues raised by each case study, the participants selected have wide-ranging roles (see table on p.76). People therefore were not selected at random; the sampling was expert-orientated. All the participants are directly connected to either MEDIN, eCrystals, LabTrove, FLLOC and SPLLOC unless stated otherwise (during sections 4.2, 5.2 and 6.2 in Chapters 4-6).

While initially undergraduate and postgraduate students understandably were found to be less experienced with such issues and therefore not selected as participants, an exception was made for the crystallographic PhD student – **Miss J.** – who has conducted extensive research in the area of data re-usage and has contributed to the development of eCrystals and LabTrove.

Each interview participant falls under five categories of main expertise: (1) data management, (2) data policy, (3) academia, (4) legal and (5) technology. This is to aid

---

[188] Alicia Wise, 'Collaborative transformations in scholarly publishing' in *International Yearbook of Library and Information Management 2004-2005: Scholarly Publishing in an Electronic Era*, ed. by G.E. Gorman (London: Facet Publishing, 2005), pp. 20-31 (p. 24).

[189] C. Gibson, 'Semi-structured and unstructured interviewing: a comparison of methodologies in research with patients following discharge from an acute psychiatric hospital', *Journal of Psychiatric and Mental Health Nursing*, 5 (6) (1998), 469-477 (p. 470) <http://dx.doi.org/10.1046/j.1365-2850.1998.560469.x>

cross-comparison by offering a consistent balance of expertise within each case study. Although it is obvious that each participant will have overlapping areas of expertise, it was vital that these five categories were covered by participants with extensive knowledge it that particular area.

Data management expertise is required, because these individuals have a comprehensive overview of the complete academic research data lifecycle from collection through to maintenance and re-usage in a given research project and/or through a particular platform. Due to the range of expertise required, two participants per case study fall under this category. Chapter 2 has further shown that data policy forms a significant part of academic best practice and socio-cultural frameworks. Therefore, individuals with extensive data policy expertise are selected to explore how data policy has impacted on each model. The academic category encompasses individuals with particular domain-specific expertise who are often data originators, managers and users. Finally, exploration of legal and technological frameworks necessitates specialist input from legal and IT professionals.

Overall, these five categories reflect the wealth of expertise required to critically discuss the impacts of provenance metadata, legal, technological and socio-cultural issues arising within each case study, and how best practice lessons can be drawn from them individually and together. Refer to the following table for a full overview of the people interviewed (their pseudonym, role, organisation and main area of expertise):

| *Cross-comparison table* | People interviewed | | |
|---|---|---|---|
| **Main area of expertise:** | **Chapter 4: MEDIN** | **Chapter 5: eCrystals and LabTrove** | **Chapter 6: FLLOC and SPLLOC** |
| **Data management** | **Mr B.**<br>MEDIN Core Member.<br>MEDIN. | **Dr G.**<br>Crystallographic Academic.<br>Chemistry Department at the University of Southampton. | **Dr O.**<br>Second Language Acquisition Academic.<br>Modern Languages Department at the University of Southampton. |
| | **Ms E.**<br>Data and Biodiversity Scientist.<br>The Data Archive for Seabed Species and Habitats. | **Dr A.**<br>Crystallographic researcher.<br>Chemistry Department at the University of Southampton. | **Dr P.**<br>Second Language Acquisition Academic.<br>Modern Languages Department at the University of Southampton. |
| **Data policy** | **Mr N.**<br>Member of External Relations.<br>The UK Hydrographic Office. | **Mr C.**<br>Data Manager within Library Services at the University of Southampton. | **Miss L.**<br>Academic Liaison Librarian.<br>Library Services at the University of Southampton. |
| **Academia** | **Dr S.**<br>Physical Oceanography Academic.<br>The National Oceanography Centre at the University of Southampton. | **Miss J.**<br>Crystallographic PhD Student.<br>Chemistry Department at the University of Southampton. | **Dr K.**<br>Interdisciplinary Academic.<br>Modern Languages Department at the University of Southampton. |
| **Legal** | **Mrs T.**<br>Intellectual Property and Licensing Officer.<br>The UK Hydrographic Office. | **Mr H.**<br>Legal Advisor.<br>Legal Services at the University of Southampton. | **Mr M.**<br>Legal Advisor.<br>Research and Innovation Services at the University of Southampton. |
| **Technology** | **Mr W.**<br>Data Scientist.<br>The British Oceanographic Data Centre. | **Ms R.**<br>Software Engineer.<br>Electronics and Computer Science Department, University of Southampton. | **Mr D.**<br>Member of the Technical Solution's Team within iSolutions at the University of Southampton. |

Table 1 People interviewed (Chapters 4-6): their pseudonyms, roles and expertise

### 3.4.2 University of Southampton: open access and open data

Thirteen interview participants were working at the University of Southampton during the interviews. It is therefore important to briefly address the academic research data re-usage landscape at the University. The University of Southampton is a Russell Group University and the case studies are therefore well-placed to raise issues that are representative of many UK higher education institutions. Moreover, since 1999 the University of Southampton has had strong connections to the open access movement.[190] Therefore, it is expected that the thirteen participants from the University should be knowledgeable about academic research data re-usage in the wider scope of the digital open movements. It is further anticipated that the other five participants not directly connected to the University – all the MEDIN case study participants apart from **Dr S.** – are also well-informed, because of their experiences with open government data policy and the Open Government Licence.

The EPrints repository software was developed at the University of Southampton to support the long-term preservation of, and access to, a range of academic research materials from pre-print and post-print journal articles to e-theses.[191] The University of Southampton has its own institutional research repository – ePrints Soton – which runs on this software.[192] Since '2006 the University mandated all research to be recorded in ePrints Soton and since 2008 all PhD and MPhil theses have also been deposited'.[193] Furthermore, this software is used by a number of institutional and non-institutional repositories around the world; from Padua@Research at the University of Padua, Italy for electronic research documents and e-theses, to Caltech THESIS at the University of California, USA for the University's engineering and

[190] Mark Brown, 'Open Access at the University of Southampton: Pushing the boundaries and the art of the possible – Case study', *JISC case study report*, Doc# 796, Version 1.1 (October 2011) <http://www.jisc.ac.uk/media/documents/topics/openaccess/JISC_SouthamptonCase_1.pdf> [accessed 9 August 2015].

[191] 'EPrints Contributors', *EPrints Website* <http://www.eprints.org/software/contributors/> [accessed 9 August 2015].

[192] *ePrints Soton Website* <http://eprints.soton.ac.uk/> [accessed 9 August 2015].

[193] 'Open Access and ePrints Soton: Introduction', *University of Southampton Website* <http://library.soton.ac.uk/openaccess> [accessed 9 August 2015].

masters e-theses.[194] Moreover, a number of academics from the University of Southampton, and GNU EPrints are signatories to the Budapest Open Access Initiative, and the University hosted the third Berlin Conference (see section 2.3.1).[195]

More recently the University established strong links to the open data movement.[196] University of Southampton academics advise the UK government on the establishment and development of data.gov.uk to provide access to open government data, and co-founded the Open Data Institute.[197] In October 2011 to March 2013, the DataPool funded by the JISC Managing Research Data Programme 2011-13 focused on 'creating services, systems and support for collecting and managing research data across the large multidisciplinary institution that is the University of Southampton.'[198]

For reasons of research transparency, it is essential to highlight the University of Southampton's contribution to the open access and open data movements. It is recognised that the University has a pro-open ethos and this may influence its staff and students. However, this is seen as advantageous as interview participants at the University are more likely to be well-informed about these issues.

---

[194] *Padua@research Website* <http://paduaresearch.cab.unipd.it/> [accessed 9 August 2015]; *Caltech THESIS Website* <http://thesis.library.caltech.edu/> [accessed 9 August 2015].

[195] 'View signatures: Southampton search'. *Budapest Open Access Initiative Website* <http://www.budapestopenaccessinitiative.org/list_signatures?indorg=all&keyword=southampton> [accessed 9 August 2015]; 'Berlin 3 Open Access: Progress in Implementing the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities', 28 February-1 March 2005, University of Southampton <http://www.eprints.org/events/berlin3/outcomes.html> [accessed 9 August 2015].

[196] The University of Southampton also facilitates access to a number of administrative open data, such as research facilities and academic programmes called University of Southampton Open Data, and maintains a hub for open data from UK academic institutions. In 2012, University of Southampton Open Data won a Times Higher Education Award for Outstanding ICT Initiative of the Year. Refer to: 'Open Data from UK Academic Institutions <http://hub.data.ac.uk/> [accessed 9 August 2015]; 'University of Southampton Open Data Service', *University of Southampton Website* <http://data.southampton.ac.uk/> [accessed 9 August 2015]; John Gill, 'Times Higher Education Awards 2012 winners', *The Times* <http://www.timeshighereducation.co.uk/story.aspx?storyCode=2003147> [accessed 9 August 2015].

[197] 'Sir Tim Berners-Lee: President and Co-Founder', *'Open Data Institute' (ODI) Website* <http://www.theodi.org/people/timbl> [accessed 9 August 2015]; 'Sir Nigel Shadbolt: Chairman and Co-Founder', *'Open Data Institute' (ODI) Website* <http://www.theodi.org/people/nrs> [accessed 9 August 2015]; Tim Berners-Lee and Nigel Shadbolt, 'Our manifesto for government data', *Guardian*, 21 January 2010 <http://www.theguardian.com/news/datablog/2010/jan/21/timbernerslee-government-data> [accessed 9 August 2015]; Tim Berners-Lee and Nigel Shadbolt, 'There's gold to be mined from all our data'. *The Times*, 31 December 2011 <http://www.thetimes.co.uk/tto/opinion/columnists/article3272618.ece> [accessed 9 August 2015].

[198] *Jisc DataPool Project Website* <http://datapool.soton.ac.uk/about/> [accessed 9 August 2015].

The case studies selected within Chapters 4-6 each have a connection to a particular thesis supervisor (as previously explained in section 3.3). However, all three case studies deliberately lie outside the thesis author's prior higher education experience to prevent pre-given assumptions concerning academic research data re-usage.

Two of the case studies – Chapter 5: eCrystals and LabTrove, and Chapter 6: FLLOC and SPLLOC – are based at the University of Southampton. However, the cross-institutional and international reach of eCrystals, LabTrove, FLLOC and SPLLOC should not be underestimated. While the everyday administration of these four platforms takes place at the University of Southampton, contributions are made from a number of external institutions.

For instance, eCrystals was founded as part of a collaborative project between the University of Bath and the University of Southampton (see Chapter 5, section 5.1.1.1 for more information). As the crystallographic repository of the National Crystallographic Service, eCrystals is reliant on crystallographic samples from researchers across the UK (and beyond – in a small number of cases). While LabTrove software was developed at the University of Southampton its use is not confined to the University; the University of Sydney utilised LabTrove as part of its Open Malaria Project.[199]

Moreover, the FLLOC website was established as part of a collaboration between the University of Newcastle and the University of Southampton (see Chapter 6, section 6.1.1 for more information). FLLOC brings together nine second language acquisition corpora, four of which were produced by other universities: Brussels Corpus, Reading Corpus, Salford Corpus and the University of East Anglia Corpus. FLLOC and SPLLOC are also subscribers to the international Child Language Data Exchange System (CHILDES) based at Carnegie Mellon University in the USA (see

[199] 'Open Malaria Project', Open Source Malaria Website <http://opensourcemalaria.org/> [accessed 9 August 2015]; *Open Malaria Project Website* <http://malaria.ourexperiment.org/> [accessed 9 August 2015]; Jeremy Burrows, 'Advancing antimalarial drug research through open source initiatives', *Guardian*, 24 July 2013 <http://www.theguardian.com/global-development-professionals-network/2013/jul/24/open-source-drug-discovery-research> [accessed 9 August 2015].

Chapter 6, section 6.1.4 for more information).[200] Furthermore, the LANGSNAP Corpus has partners in France, Spain and Mexico (see Chapter 6, section 6.3.4.6).

In summary, these case studies – although seemingly local – were chosen to represent not only UK Russell group universities, but to further reflect the cross-institutional (and global) realities of academic research data re-usage. Although principally based at the University of Southampton, eCrystals, LabTrove, FLLOC and SPLLOC are products of a number of higher education institutions and model the complexity of research projects in other domains.

### 3.4.3 Use of the interviews

At the heart of this thesis is the principle of good academic conduct, research practice and their extension. From the outset, this thesis signed up to the University of Southampton's code of practice and research ethics.[201] The semi-structured interviews were approved by the University of Southampton's Management Ethics Committee: Paper System in August 2011 with further additions approved in January 2012.

> **It must be noted that all views are the participants' own they may not represent the views of the Marine Environmental Data and Information Network (MEDIN), the British Oceanographic Data Centre (BODC), The Data Archive for Seabed Species and Habitats (DASSH), the UK Hydrographic Office (UKHO), eCrystals, LabTrove, Research and Innovation Services (RIS) at the University of Southampton, the Library at the University of Southampton, Legal Services at the University of Southampton, the French Learner Language Oral Corpora (FLLOC), the Spanish Learner Language Oral Corpora (SPLLOC), LANGSNAP or the University of Southampton. Any inferences drawn from these interviews belong to the thesis author and may not necessarily be held by participants or their respective organisations.**[202] (Bold emphasis in original documentation.)

The thesis respects and upholds the confidentiality of each interview participant. Interview participants are given an anonymous pseudonym and a random gender to

---

[200] Brian MacWhinney, 'The CHILDES Project: Tools for Analyzing Talk – Part 1: The CHAT Transcription Format', *CHILDES Manual*, Carnegie Mellon University (19 August 2015) <http://childes.talkbank.org/manuals/CHAT.pdf> [accessed 9 August 2015].

[201] 'University of Southampton Ethics Policy', *University of Southampton Website* <http://www.southampton.ac.uk/inf/ethics_policy.html> [accessed 9 August 2015].

[202] This notice is required under part 1.4 of Laura German's 'Semi-Structured Interviews: Participation Information Document' Approved by the Management Ethics Committee August 2011 (further additions approved in January 2012.) – Paper System. Refer to Appendix A.2 for the full document.

maintain confidentiality. All the interview participants' data are made confidential in accordance with the approved ethics application entitled: Semi-Structured Interviews – Participation Information Document (see Appendix A.2). This document includes: the semi-structured interview procedure; participant consent information; best practice; and, consent form. Participants were asked to inspect this document. They received a digital copy in advance of the interview, and a paper copy with the designed semi-structured interview questions attached at the interview.

All interview participants initialled and signed a consent form. Permission was given by all participants for full sound recordings of the interviews to be copied to a CD which was appended to this thesis. This thesis is unable to offer full anonymity however. Participants are informed that disclosure of their roles and experience is crucial to the objectives of this thesis.

### 3.4.4 Data accessibility

The sounding recordings (raw data) were only made available to Laura German (thesis author), supervisors, examiners and interview participants (where requested). These sound recordings enabled supervisors and examiners to check the validity, accuracy and impartiality of the semi-structured interview data, analysis and interpretation.

Any copies of the sound recordings were deleted after final submission of the thesis.[203] However, as roles, quotes and other labelled insights (modified data) are an integral part of this thesis, permission has been received from all participants to potentially openly release this (e-)thesis in the future. The participants all initialled the following statements on the consent form:

> 5. The participant is made aware that the University of Southampton mandates that all its theses are made openly accessible in its repository on the Web.
> 6. The participant is made aware that the researcher may wish to publish this research in the future.[204]

It was decided during the early stages of the PhD that open access to full verbatim transcripts and sound recordings were potential disincentives to interview participation. For instance, Barbara DiCicco-Bloom and Benjamin F. Crabtree contend: 'tape-recorded data can be a source of danger for those who are taped because recorded data

[203] iSolutions were asked (via email) whether simply deleting the interview files from the Dictaphone was sufficient, it was confirmed that there were no further actions to take.
[204] 'Semi-Structured Interviews: Participation Information Document', p. 9.

is incontrovertible'.[205] A number of participants stated that they would not have taken part in this research but for the high level of confidentiality stipulated. This is a point raised by **Dr K.** in Chapter 6 who admitted that despite her role as an interviewer in the field of life histories research, she was apprehensive about being interviewed until she read all the interview documentation, and was able to acknowledge the anonymisation procedures used [$K_{24}$].[206] This is because she realises: 'how easy it is for digital data to be transmitted *um* and used in all sorts of different ways nowadays' [$K_{24}$].

It has been important to keep in contact with interview participants and offer updates about publications in conference proceedings. Individuals from Legal Services at the University of Southampton, MEDIN, eCrystals, LabTrove, FLLOC and SPLLOC have also been asked for their permission to use screen shots during Chapters 4-6.

### 3.4.5 Semi-structured interview questions

A consistent approach to each interview enables reliable cross-interview data analysis and interpretation. All participants were interviewed in person and individually at their organisation and location of choice, where interviews were recorded by Dictaphone.[207] Alongside a range of diverse, unplanned questions, each interview participant was asked at least nine designed semi-structured interview questions (see Table 2 on the next page) – unless stated otherwise in the interview record table (due to human error or where a question was already covered by the participant). Participants gave their consent to be asked further questions via email if required to cover any new points that arose after the interviews and as the thesis developed.

Question 3) was added after the MEDIN interviews were completed, and asked to all subsequent interview participants. This was due to the prominence of quality

---

[205] Barbara DiCicco-Bloom and Benjamin F. Crabtree, 'The qualitative research interview', *Medical Education,* 40 (4) (2006), 314-321 (p. 318) <http://dx.doi.org/10.1111/j.1365-2929.2006.02418.x>

[206] See section 3.4.7 for a detailed explanation about the labelling system used – i.e. [$K_{24}$].

[207] For information about the 'advantages and disadvantages of conducting interviews by email, telephone, instant messenger or face-to-face' refer to: Raymond Opdenakker, 'Advantages and Disadvantages of Four Interview Techniques in Qualitative Research', *Forum: Qualitative Social Research,* 7 (4) (2006), 1-13 <http://www.qualitative-research.net/index.php/fqs/article/view/175/392> [accessed 9 August 2015]; for information about selecting an appropriate interview location refer to: Sarah A. Elwood and Deborah G. Martin, '"Placing" interviews: location and scales of power in qualitative research', *Professional Geographer,* 52 (4) (2000), 649-57 (p. 651) <http://www.utsc.utoronto.ca/~kmacd/IDSC10/Readings/interviews/place.pdf> [accessed 9 August 2015].

assurance and responsibility issues raised within the MEDIN interviews, which led the thesis to investigate this important area in more detail. This reflects the intellectual development of research, as it organically changes over the course of the thesis. Moreover, any preconceptions the researcher has are almost invariably challenged.

| **Table of Designed Semi-Structured Interview Questions** | |
|---|---|
| ***'X' = The interview participant's organisation. For example, replace 'X' with MEDIN, Legal Services, eCrystals and FLLOC.*** | |
| **Question 1).** | What is your role at *X* and why do you work here? |
| **Question 2).** | i) How does *X* make academic research data reusability more effective and appropriate? Please give examples. |
| | ii) And are there procedures, systems or software issues which help/hinder effective research data reusability? |
| **Question 3). [Added to designed semi-structured interview questions after MEDIN interviews were conducted.]** | i) What is the quality assurance process within *X*? |
| | ii) What is your role and responsibility with regard to assuring the quality of data? |
| | iii) Does this responsibility differ between data authors, data managers and data users? |
| **Question 4).** | i) What is *X*'s role within scholarly communication? |
| | ii) Does *X* impact on publishers, funding bodies and academics? If so how? |
| **Question 5).** | i) What are the main legal issues that impact on data reusability at *X*? |
| | ii) Have there been any examples of misuse of data? |
| **Question 6).** | How important is provenance metadata at *X*? What is its function? |
| **Question 7).** | What other external issues arise that have an impact on data at *X*? For example, ethical issues, sensitive data, data protection, use of data for commercial purposes? |
| **Question 8).** | With regard to more effective data reusability, where do you see *X* in the next five years? The next twenty years? What is most likely to influence such developments? |
| **Question 9). [Case Study Specific.]** | MEDIN) Conservation and exploitation of marine resources seem to have contradictory research drivers. Does MEDIN intervene to maintain a representative balancing of these drivers? |
| | eCrystals and LabTrove) To what extent is there a culture of sharing within chemistry? How has the web improved access to chemistry data? How does this impact on assuring these data as reliable, robust and fit for purpose? |
| | FLLOC, SPLLOC and LANGSNAP) Oral histories/second language acquisition recordings are personal academic research data – how does this human element impact on the facilitation of transparent research methodologies and quality assurance in contrast to non-personal academic research data? How are ethics independently, quality assessed? How has the web improved access to Oral histories/second language acquisition research data? |
| **Question 10)** | Do you have any further points that you would like to raise and discuss? |

Table 2 Table of ten designed semi-structured interview questions

Question 1) is an icebreaker question 'to ease the participant into the interview and key questions relating to the study objectives.'[208] Question 10) is similar, as it eases the participant out of the interview and presents an opportunity for the participant to add any relevant information that may not have been discussed or even considered by the researcher. Questions 2) to 7) are grouped into issues around: 2) how the data platform works and technological issues (technological framework and issues); 3) socio-cultural issues – quality assurance, responsibilities, and ethics issues; 4) how the data platform compares to the traditional academic publication model; 5) legal framework and issues; 6) provenance framework and issues; 7) any other issues not covered; and, 8) potential future developments for the data platform. To enable direct examination of exclusive features particular to each data platform, Question 9) is a case study specific question.

Three types of question – descriptive, normative and cause-and-effect – were utilised in the design, as they encourage a wide range of response during interviews.[209] While designing these ten questions, it was important to avoid systematic bias, such as closed questions with only yes and no answers, leading questions with bias, multiple questions, false alternatives and rhetorical questions.[210] The questions were written in plain language without specialist terminology, as the participants come from wide-ranging backgrounds where such terms are not familiar or convey different meanings within disciplines.[211]

### 3.4.6 Interview analysis and interpretation

For research transparency and scrutiny, it is important that data analysis and interpretation is also explained, as Michael Quinn Patton states:

---

[208] McCance and others, p. 351.

[209] For further information about descriptive, normative and cause-and-effect questions refer to: Linda G. Morra-Imas and Ray C. Rist, *The Road to Results: Designing and Conducting Effective Development Evaluations* (Washington DC, USA: The International Bank for Reconstruction and Development/The World Bank, 2009) pp. 223-229. Google eBook.

[210] Peter McIlveen and others, 'Evaluation of a semi-structured assessment interview derived from systems theory framework', *Australian Journal of Career Development,* 12 (3) (2003), 33-41 (p.13) <http://dx.doi.org/10.1177/103841620301200306>

[211] Morra-Imas and others, p. 222.

> However analysis is done**, analysts have an obligation to monitor and report their own analytical procedures and processes as fully and truthfully as possible**.[212] (Patton's original bold emphasis.)

Catherine Pope and others state that: 'these data are not necessarily small scale: transcribing a typical single interview takes several hours and can generate 20-40 pages of single spaced text.'[213] Raymond Lee and Nigel Fielding label transcription as 'a major bottleneck', as it may take on average four to eight hours to transcribe an hour's worth of sound recording.[214] The eighteen semi-structured interviews generated around fourteen hours of interview sound recordings. For the purpose of this research, the production of full verbatim transcriptions warrants an uneconomical use of time and excessive, resulting material. Graham Gibbs contends: 'you are more likely to focus on the bigger picture and not get bogged down in the details of what people have said.'[215]

Transcriptions are only a descriptive record not an analysis and interpretation.[216] The production of full verbatim transcriptions is not a necessity for all interview analyses, particularly where methodologies find no substantial benefit in their employment.[217] For example, this thesis is not interested in examining the interviews for their linguistic or grammatical fluency and hence analysing them by means of discourse and conversation analysis.[218] The qualitative data from the semi-structured interviews were directly analysed from the sound recordings, and analysis software that requires full verbatim transcriptions were not utilised.

---

[212] Michael Quinn Patton, *Qualitative Research and Evaluation Methods*, 3rd edn (London: SAGE Publications, 2002), p. 434.

[213] Catherine Pope, Sue Ziebland and Nicholas Mays, 'Analysing qualitative data', *British Medical Journal*, 320 (7227) (2000), 114-116 (p.114) <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1117368/> [accessed 9 August 2015].

[214] Raymond M. Lee and Nigel G. Fielding, 'Chapter 23: Tools for Qualitative Data Analysis' in *Handbook of Data Analysis*, ed. by Melissa Hardy and Alan Bryman (London: SAGE Publications, 2004), pp. 529-546 (p. 533). Google eBook.

[215] Graham Gibbs, *Analyzing Qualitative Data* (London: SAGE Publications, 2007), p. 11. Google eBook.

[216] Pope and others, p. 114.

[217] Gibbs, p. 11.

[218] Discourse analysis is defined by three parts – (1) 'grammar beyond the sentence', (2) 'language in use' and (3) the 'rhetoric of power' – which are not relevant to this thesis. For further information refer to: Harvey Russell Bernard and Gery W. Ryan, *Analyzing Qualitative Data: System: Systematic Processes* (London: SAGE Publications, 2010), p. 222. Google eBook.

There is no single method to analyse and interpret data collected by semi-structured interviews.[219] This is perhaps due to the variety of disciplines that employ this research method, from human geographers and health care professionals to computer scientists.[220] There are three main approaches to qualitative data analysis: (1) the quasi-statistical approach; (2) the thematic coding approach; and, (3) the grounded theory approach.[221] The quasi-statistical approach transforms qualitative data into quantitative data, such as obtaining word or phrase frequencies.[222] This thesis finds the quasi-statistical approach unwarrantable, as it is highly likely that legal experts will use legal terms more often than non-legal experts. This approach is better suited to a greater number of individuals interviewed with similar roles. In addition, this thesis does not produce full verbatim transcripts; therefore, the available software to retrieve text, word frequency extrapolation, or code transcriptions is impracticable.[223] The thematic coding approach labels data using codes to organise data with the same label into themes which 'are then compared and contrasted with any similar material in other sources'.[224] The grounded theory approach is a type of thematic coding that is 'used to develop a theory 'grounded' in the data.'[225]

---

[219] Robert K. Yin, *Case Study Research: Design and Methods*, 4th edn (London: SAGE Publications, 2009), pp. 162; Catherine Marshall and Gretchen B. Rossman, *Designing Qualitative Research*, 4th edn (London: SAGE Publications, 2006), p. 154; Matthew B. Miles, 'Qualitative Data as an Attractive Nuisance: The Problem of Analysis', *Administrative Science Quarterly*, 24 (4) (1979), 590-601 <http://dx.doi.org/10.2307/2392365>

[220] For examples refer to the following articles: Mike Crang, 'Qualitative methods: the new orthodoxy?', *Progress in Human Geography*, 26 (5) (2002), 647-655 <http://dx.doi.org/10.1191/0309132502ph392pr>; Suzanne Moffatt and others, 'Using quantitative and qualitative data in health services research – what happens when mixed method findings conflict?', *BMC Health Services Research*, 6 (28) (2006), 1-10 <http://dx.doi.org/ 10.1186/1472-6963-6-28>; Larry E. Wood, 'Semi-Structured Interviewing for User-Centered Design', *Interactions*, 4 (2) (1997), 48-61 <http://dx.doi.org/10.1145/245129.245134>

[221] Jeremy Jolley, *Introducing Research and Evidence-Based Practice for Nursing and Healthcare Professionals*, 2nd edn (Oxford: Routledge, 2013), p. 204. Google eBook; For information regarding the pros and cons of analysis software refer to: Colin Robson, *Real World Research*, 3rd edn (Chichester: Wiley, 2011), p. 470-473.

[222] Robson, p. 467.

[223] Lee and Fielding, p. 531.

[224] Lee and Fielding, p. 530.

[225] Barney G. Glaser and Anselm L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*, 7th paperback printing (New Jersey, USA: Transaction Publishers, 2012). Google eBook.

This thesis employs the thematic coding approach. Each interview had three written outputs produced after each interview had taken place: (1) initial write-up, (2) a second write-up, and (3) a third write-up. The initial write-up paraphrased answers and compiled a table which recorded the question number or the full non-designed question, and the timings of a question and its answer. The second write-up selected and transcribed generally important quotes that directly addressed the research questions or raised a pertinent grey area. This resulted in a mixture of paraphrase, interpretation, quotes and provenance metadata. Under the third write-up the data were split into loose sub-categories, including: model background and aims, provenance, quality assurance, permissions, acknowledgement, and formats. These were then assigned to four overarching categories – provenance metadata issues, legal issues, technological issues and socio-cultural issues.

There was cross-comparison between the third write-ups gathered from the six interview participants per case study. Data were cross-compared under the corresponding loose categories without direct comparison of answers to the same question. This was because the loose categories not only captured the same answers to different questions, but the associated questions that arose organically within each interview. Chapters 7-8 offer a further level of cross-comparison scrutinising the similarities and differences that are raised by each case study.

### 3.4.7 Interview records and traceability

Each interview has a table that records the interview for every participant, containing the exact questions asked and their duration. In the event that a question is not asked within the interview or directly referenced within the thesis a justification is given; all quotations and paraphrases match original recordings. These tables are available in Appendix B. This is for research transparency, and to enable examiners and supervisors to locate quotations and paraphrases on the sound recordings with ease. To showcase immediacy and transparency, verbatim quotations are not edited. Natural utterances that denote hesitation and pauses within speech are included, such as *um* and *er*. The inclusion of natural utterances is standard practice for academics working within linguistics, because the hesitations and reformulations may reveal further information.

Every direct reference to a specific question in Chapters 4-6 has a table reference at the end of that particular sentence. The table reference encompasses the letter from

the pseudonym and a section number refers to a specific question asked to the participant that is located within the record table in the appendix. For example: '**Mr Y.** maintains that the Web has changed the ways in which data are re-used [$Y_7$].' This label [$Y_7$] corresponds to the 7th question asked to **Mr Y.** (not the 7th designed question). This labelling is vital for traceability by ensuring that all views and quotes are correctly attributed to each participant. For thesis supervisors and examiners, it was also a useful means of finding a specific question and answer on the sound recording.

It required a considerable amount of time and effort to record and label all the interview data in this manner. However, it is to ensure with reasonable confidence people who are authorised to access the sound recordings (thesis author, supervisors and examiners) could get the attribution information out of the data quickly and efficiently and find answers to specific questions with ease. As this thesis is part of the first cohort of web science research, there was no methodological precedent. The final labelling system was produced through trial and error. It developed from a footnote system that recorded the pseudonym of the participant, time of interview and question asked. This initial system was too repetitive and lacked immediacy. Therefore, it was changed to the simple reference note system, which places the participant's pseudonym letter and interview section number in square brackets within the main body of the thesis (for example [$D_7$]).

## 3.5 Methodology: summary

In summary, Chapters 4-6 focus on the three case studies – Chapter 4: MEDIN, Chapter 5: eCrystals and LabTrove, and Chapter 6: FLLOC and SPLLOC. To facilitate cross-comparative evaluation within Chapters 7-8, Chapters 4-6 follow a similar structure. Each chapter is divided into four parts.

The first part presents the subject of the chapter (MEDIN, eCrystals and LabTrove or FLLOC and SPLLOC respectively) of the case study by focusing on its primary source materials. The primary source materials largely comprise of information found on the subject's website and grey literature. This initial part therefore evaluates the rationale, legal framework, technological framework and socio-cultural framework of the case study. It thus raises the issues that require additional consideration during the semi-structured interviews. It further outlines which secondary issue, raised by Hwang and others, the chapter confronts. Moreover, permission was received (via

email) for website screen shots of each model to be provided within this thesis for purposes of illustration (subject to attribution).

The second part of the chapter then provides information about the people interviewed, including their pseudonyms, roles and connections with the case study. There is a further reminder about the interview tables within the appendices.

The third part of the chapter explores the interpretation and analysis of the semi-structured interview materials. Here the materials are examined critically under four sub-headings: provenance metadata issues; legal issues; technological issues; and, socio-cultural issues.

Interim conclusions evaluate the extent to which the model explored in the case study resolves the issue raised by the case of Hwang and others. In consequence, it further assesses whether the model makes for best academic practice. Finally, each chapter outlines the key themes and grey areas to be re-visited during Chapter 7: Recommendations and Chapter 8: Conclusion and Future Work.

# Chapter 4: MEDIN Case Study

This case study confronts the first major issue of Hwang and others: how diverse types of academic research data from multiple originators, contexts and sources can be safeguarded for a wide user base. To explore this issue an existing UK model of data re-usage, centred on provenance metadata within the marine sciences, the Marine Environmental Data and Information Network (MEDIN) has been selected:[226]

> The MEDIN portal is a metadata discovery service providing users with a single point of access to a well-balanced, authoritative marine metadata catalogue. […] Metadata records are available for UK marine data sets across all subject areas and disciplines.[227]

The MEDIN model is built around a robust provenance metadata framework called discovery metadata, which enables research users to locate particular marine environmental data across a number of accredited sources.[228]

Since 2008, MEDIN has safeguarded marine environmental data from multiple sectors in the UK, including businesses, universities, private contractors, government departments and private bodies. While its wide-ranging role is acknowledged, this case study concentrates on its position as a guarantor and disseminator of academic research data in a digital age.

MEDIN has seven accredited thematic data archive centres based across the UK: (1) the British Geological Survey (BGS) which manages seabed and sub-seabed geology, geophysics data; (2) the British Oceanographic Data Centre (BODC) which maintains water column oceanographic data; (3) the Data Archive for Marine Species and Habitats Data (DASSH) which safeguards biodiversity data; (4) the UK Hydrographic Office (UKHO) which curates bathymetry data; (5) the MET Office which preserves meteorological (metocean) data; (6) the Archaeology Data Service which safeguards marine historic environment data; and, (7) the FishDAC containing fisheries data that are hosted by the Centre for Environment, Fisheries and Aquaculture

---

[226] *The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/> [accessed 9 August 2015].

[227] 'Search the MEDIN Data Archive Centres', *The Marine Environmental Data and Information Network (MEDIN) Website* <http://portal.oceannet.org/search/full> [accessed 9 August 2015].

[228] 'Publishing', *The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/marine_data_standards/other_marine_data_standards/what_is_metadata_index.html> [accessed 9 August 2015].

Science (Cefas), Marine Scotland and DASSH.[229] MEDIN has to manage a number of different licensing regimes which are in operation across its thematic data archives centres, including the Open Government Licence and Creative Commons licences (as briefly introduced in Chapter 2, section 2.3.3).

In 2012, during the MEDIN interviews, there were only four accredited data archive centres, namely: BGS, BODC, DASSH and UKHO. Alongside an interview with **Mr B.** who is a MEDIN core member and **Dr S.** a physical oceanography academic, interviews were conducted with members of BODC, DASSH and the UK Hydrographic Office only.

While it must be highlighted that the UK Hydrographic Office's data archive for bathymetry data was accredited but not yet functioning at the time of the interviews, this shows the willingness of **Mrs T.** and **Mr N.** to offer their insights before these data were made available. Together with examining the overarching provenance metadata, legal, technological and socio-cultural frameworks employed by MEDIN, this case study further investigates three of the original MEDIN data archive centres.

This case study begins by addressing the interdependent legal, technological and socio-cultural frameworks that underpin MEDIN through assessment of its primary source materials – found on the MEDIN website and the websites of its accredited data archive centres. It then moves on to critical examination of the semi-structured interview materials, before evaluating to what extent MEDIN resolves the first, major issue raised by Hwang and others.

[229] 'Submitting data', *The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/data_submission/index.html> [accessed 9 August 2015].

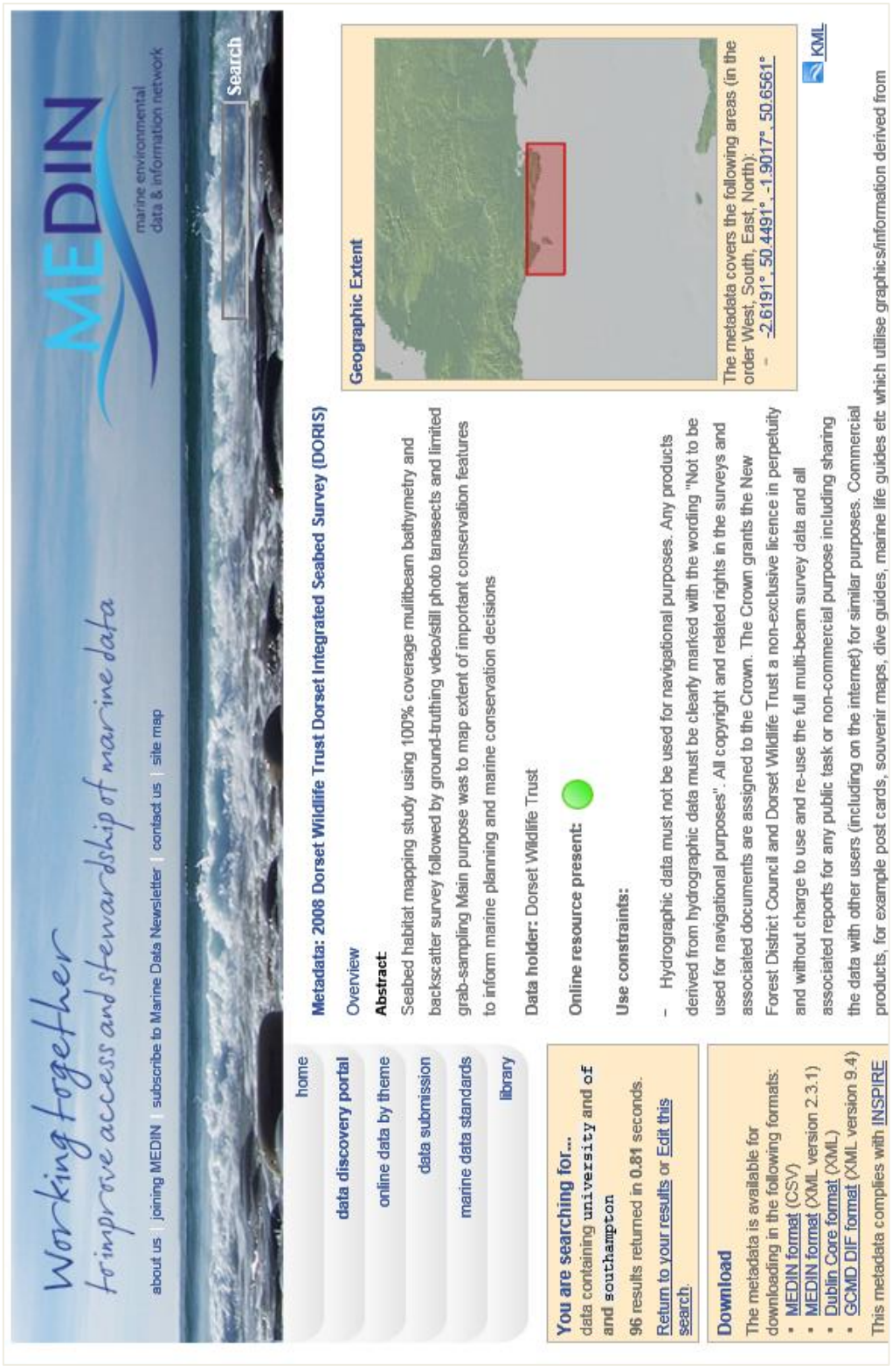
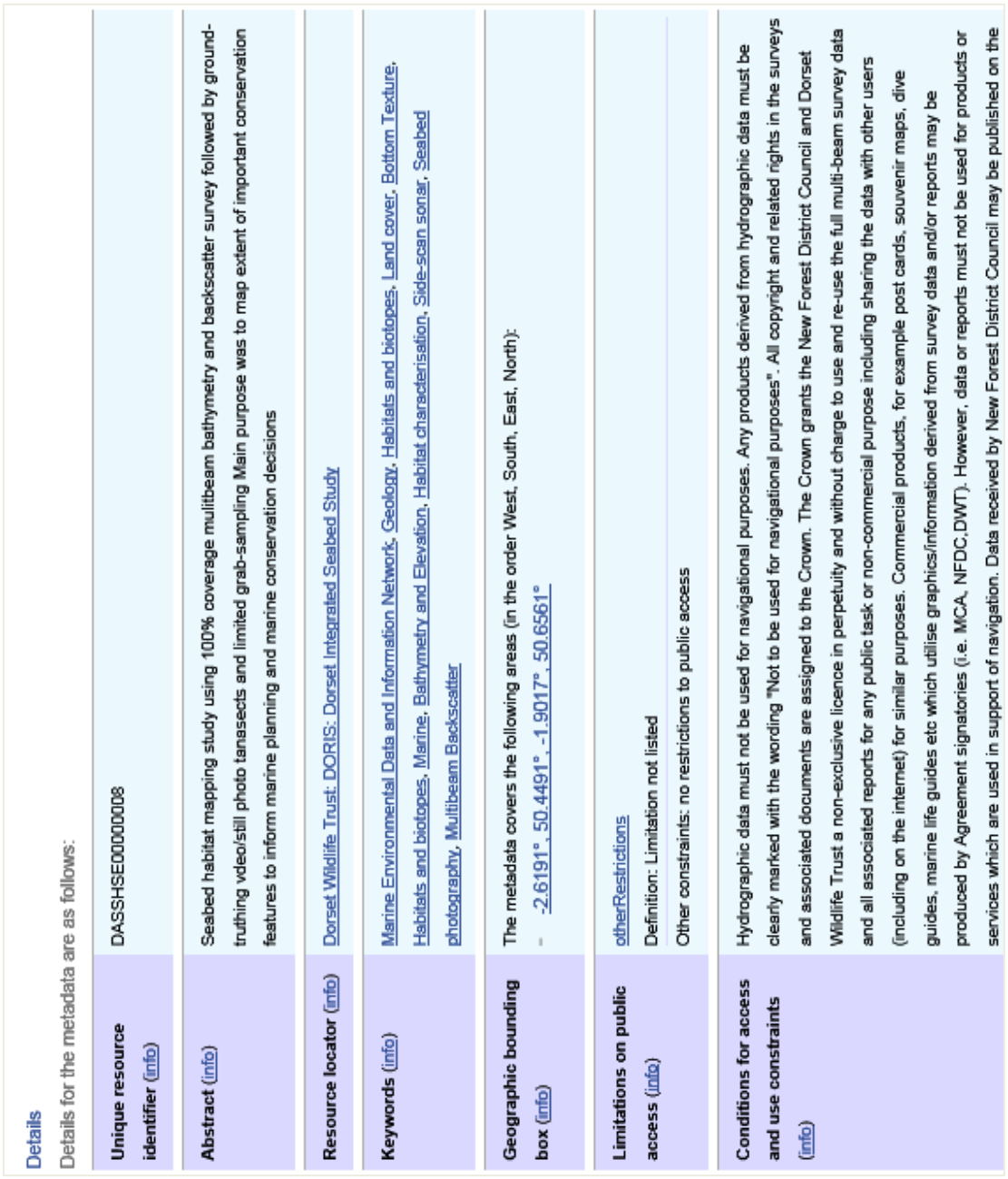



**Figure 1 Screen Shot A of 'Metadata: 2008 Dorset Wildlife Trust Dorset Integrated Seabed Survey (DORIS)', *MEDIN Website* <http://portal.oceannet.org/search/full/catalogue/dassh.ac.uk__MEDIN_2.3__f3204bb9caebe79523e45327f7b7f6e6.xml> [accessed 2 November 2014].**

**Citation: MEDIN Core Team <http://www.oceannet.org/> [accessed 2 November 2014] and the GeoData Institute University of Southampton <http://www.geodata.soton.ac.uk/geodata/> [accessed 2 November 2014]. Image taken on 2 November 2014. Reproduced with permission from the MEDIN Core Team.**



Details

Details for the metadata are as follows:

| | |
|---|---|
| **Unique resource identifier** (info) | DASSHSEE00000008 |
| **Abstract** (info) | Seabed habitat mapping study using 100% coverage mulitbeam bathymetry and backscatter survey followed by ground-truthing vdeo/still photo tanasects and limited grab-sampling Main purpose was to map extent of important conservation features to inform marine planning and marine conservation decisions |
| **Resource locator** (info) | Dorset Wildlife Trust: DORIS: Dorset Integrated Seabed Study |
| **Keywords** (info) | Marine Environmental Data and Information Network, Geology, Habitats and biotopes, Land cover, Bottom Texture, Habitats and biotopes, Marine, Bathymetry and Elevation, Habitat characterisation, Side-scan sonar, Seabed photography, Multibeam Backscatter |
| **Geographic bounding box** (info) | The metadata covers the following areas (in the order West, South, East, North):<br>- -2.6191°, 50.4491°, -1.9017°, 50.6561° |
| **Limitations on public access** (info) | otherRestrictions<br>Definition: Limitation not listed<br>Other constraints: no restrictions to public access |
| **Conditions for access and use constraints** (info) | Hydrographic data must not be used for navigational purposes. Any products derived from hydrographic data must be clearly marked with the wording "Not to be used for navigational purposes". All copyright and related rights in the surveys and associated documents are assigned to the Crown. The Crown grants the New Forest District Council and Dorset Wildlife Trust a non-exclusive licence in perpetuity and without charge to use and re-use the full multi-beam survey data and all associated reports for any public task or non-commercial purpose including sharing the data with other users (including on the internet) for similar purposes. Commercial products, for example post cards, souvenir maps, dive guides, marine life guides etc which utilise graphics/information derived from survey data and/or reports may be produced by Agreement signatories (i.e. MCA, NFDC,DWT). However, data or reports must not be used for products or services which are used in support of navigation. Data received by New Forest District Council may be published on the |

**Figure 2 Screen Shot B of 'Metadata: 2008 Dorset Wildlife Trust Dorset Integrated Seabed Survey (DORIS)', *MEDIN Website* <http://portal.oceannet.org/search/full/catalogue/dassh.ac.uk__MEDIN_2.3__f3204bb9caebe79523e45327f7b7f6e6.xml> [accessed 2 November 2014].**

**Citation: MEDIN Core Team <http://www.oceannet.org/> [accessed 2 November 2014] and the GeoData Institute University of Southampton <http://www.geodata.soton.ac.uk/geodata/> [accessed 2 November 2014]. Image taken on 2 November 2014. Reproduced with permission from the MEDIN Core Team.**

## 4.1 MEDIN: primary source materials

### 4.1.1 MEDIN's rationale

The MEDIN data discovery portal was developed by the GeoData Institute at the University of Southampton, and is accessible through the MEDIN website to anyone with a Web connection.[230] To that extent it is open access. MEDIN is hosted by the British Oceanographic Data Centre (BODC) at the National Oceanography Centre in Liverpool, and is managed by a five member core team.[231] Alongside the seven accredited thematic data archive centres, MEDIN is sponsored by fifteen organisations (correct on 5 July 2015), including the Natural Environmental Research Council (NERC) that funds the majority of marine sciences research conducted by UK higher education institutions.[232]

During 2012-3, there were nearly sixty organisations actively involved with MEDIN, including the Geodata: Consultancy based at University of Southampton.[233]

---

[230] 'Search the MEDIN data archive centres', *MEDIN Website*.

[231] 'Contact us', *The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/contact_us/> [accessed 9 August 2015].

[232] 'Joining MEDIN', *The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/joining_medin/> [accessed 9 August 2015]; MEDIN Annual Report (2012-13), *'The Marine Environmental Data and Information Network (MEDIN) Website*, p. 5 <http://www.oceannet.org/library/key_documents/documents/medin_annual_report_201213_final.pdf> [accessed 5 July 2015] – list of MEDIN sponsors: 'Department of Environment Food and Rural Affairs' (DEFRA), the Scottish Government, 'Department of Energy and Climate Change' (DECC), the Met Office, Countryside Council for Wales, the Environment Agency, the Marine Management Organisation, the Maritime and Coastguard Agency, the Crown Estate, the UK Hydrographic Office, HR Wallingford, the Joint Nature Conservation Committee and the Northern Ireland Environment Agency / Agri-Food Biosciences Institute.

[233] MEDIN Annual Report (2012-13), *The Marine Environmental Data and Information Network (MEDIN) Website*, pp. 18-19 <http://www.oceannet.org/library/key_documents/documents/medin_annual_report_201213_final.pdf> [accessed 9 August 2015] – list of organisations actively involved with MEDIN in 2012-2013:

Archaeological Data Services, Agri-Food and Biosciences Institute, Atkins Global, British Geological Survey, British Oceanographic Data Centre, Centre for Environment Fisheries and Aquaculture Science, The Crown Estate, Data Archive for Seabed Species and Habitats, Department of Energy and Climate Change, Department for Environment Food and Rural Affairs, Environment Agency, EDINA: Unit of Edinburgh University, English Heritage, Finding Sanctuary, Fugro Geos, Gardline Group, Geodata: Consultancy based at University of Southampton, Historic Scotland, HR Wallingford, Inshore Fisheries and Conservation Authorities, Institute for Marine Science and Technology, JohnPepper Consultancy, Joint Nature Conservation Committee, Mainstream Renewable Power, Marine Atlas, Marine Conservation Society, Marine Management Organisation, Marine Planning Consultants, Marine Scotland Science, Marine Biological Association, Maritime and Coastguard Agency, Marine Ecological Surveys, Met Office, Ministry of Defence, Natural England, Natural Resources Wales, Natural Environment Research Council, The Northern Ireland Environment Agency, OceanWise Ltd, Ordnance Survey, Royal

MEDIN encourages any interested individual or organisation to become a sponsor or partner:

> […] It is an open partnership and its partners represent government departments, research institutions and private companies. [/] Marine data are expensive to collect and always unique in relation to time and geographical position. There are wide benefits to be gained from working together to share and properly manage these data.[234]

In 2008, MEDIN was established by the UK government as a mechanism to address insufficient data re-usage practices across the marine environmental community:

> The fundamental problem that MEDIN was established to tackle was that enormous amounts of data were being collected but in practice very little of this was available for reuse. There were over a hundred different holders of marine environmental data, with little or no coordination of standards and formats. This meant that discovering and accessing data was very difficult, and that even when sourced, the data were often unusable because of inconsistencies in standards and formats.[235]

Therefore, the principal rationale behind MEDIN is to build a re-usable and sustainable resource of quality marine sciences data. These data can be shared within the marine sciences community – particularly by researchers and policy makers – through the employment of common standards, formats and practices that have been agreed and adopted by the marine sciences community.

#### 4.1.2 British Oceanographic Data Centre (BODC): overview

Founded in 1969 by the NERC, the BODC is the UK's National Oceanographic Data Centre.[236] BODC's aims are clearly stated on its website:

> We deal with **biological, chemical, physical** and **geophysical** data. Our databases contain measurements of nearly 22,000 different variables. Many of our staff have direct experience of marine data collection and analysis. They

---

Commission on the Ancient and Historic Monuments of Scotland, Royal Commission on the Ancient and Historic Monuments of Wales , RES Offshore, Scottish Association for Marine Science, Senergy, Scottish Government, Scottish Natural Heritage, SeaZone, SETech, Scottish Environment Protection Agency, Sustainable Scotland Marine Environment Initiative, Titan Surveys, University of the Highlands and Islands, UK Hydrographic Office, and Wessex Archaeology.

[234] 'About us', *The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/about_us/> [accessed 9 August 2015].

[235] The MEDIN Executive Team, 'MEDIN Data Archive Centre Network – A Review of Future Funding Options', Online Report, MEDIN website (November 2010), 1-27 (p. 2), *The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/library/work_stream_documents/> [accessed 9 August 2015].

[236] 'Our history', *The British Oceanographic Data Centre (BODC) Website* <http://www.bodc.ac.uk/about/our_history/> [accessed 9 August 2015].

> work alongside information technology specialists to ensure that data are documented and stored for current and future use.[237] [BODC's bold emphasis.]

As the majority of academic research data produced at UK higher education institutions are funded by NERC, the data held by the BODC are subject to the NERC data policy and the NERC policy on licensing and charging for environmental data and products. Where possible the BODC data holdings are released as open data to meet the requirements of the NERC data policy as follows:

> The Data Policy details our commitment to support the long-term management of environmental data and also outlines the roles and responsibilities of all those involved in collecting and managing environmental data. Central to the policy is that NERC-funded scientists must make their data openly available within two years of collection and deposit it in a NERC data centre for long term preservation. The aim is that all NERC-funded data are managed and made available for the long-term for anybody to use without any restrictions.[238]

All academic research data funded by NERC require a data licence agreement, as stipulated by point 5 of the NERC Data Policy:

> All environmental data made available by the NERC Environmental Data Centres will be accompanied by a data licence. Data originally provided to NERC by a third-party may have their own access and licence conditions which restrict how or when we can make data available to others, in which case our data licence conditions will reflect these.[239]

Point 4 of the NERC Policy on Licensing and Charging for Environmental Data and Products provides five key reasons why licensing of environmental data is important:

> a) Ensure that Environmental Data and Information Products are safe, remain secure, are managed effectively and used appropriately;
> b) Assist NERC in managing its intellectual property rights and ensuring that due copyright acknowledgement and relevant citation is given;
> c) Protect third party intellectual property rights and meet NERC's contractual obligations;
> d) Explain NERC's limits of liability for all Environmental Data and Information Products it supplies;
> e) Ensure that where NERC's Information Products are intended to be re-used, for example within an organisation's business, or in a value-added application or

---

[237] 'What is BODC?' *The British Oceanographic Data Centre (BODC) Website* <http://www.bodc.ac.uk/about/what_is_bodc/> [accessed 9 August 2015].

[238] 'Data Policy', *NERC Website* <http://www.nerc.ac.uk/research/sites/data/policy/> [accessed 9 August 2015].

[239] 'NERC Data Policy', (p. 2.) *Natural Environmental Research Council (NERC) Website* <http://www.nerc.ac.uk/research/sites/data/policy/data-policy.pdf> [accessed 9 August 2015].

> product, appropriate terms and conditions are described, including any usage or royalty charges that may apply.[240]

Although it is clear from point 4 of the NERC policy (on licensing and charging for environmental data and products) that NERC offer an assortment of licences, there do not appear to be any examples provided on the NERC or BODC websites. Therefore, all data held have data licences devised by the BODC, which are aligned with NERC policy.[241] The BODC provide an Enquiries Officer to help research users negotiate further access to data, or data not available via web delivery.[242]

Unlike the MEDIN Data Discovery Portal and the other data archive centres, BODC requires research users to register their details and data preferences (including access to Conductivity, Temperature, Depth (CTD) profiles, current meter series and wave data series) to access some datasets.[243]

#### 4.1.3 Database for Marine Species and Habitats Data (DASSH): overview

Founded in 2005, DASSH receives core funding from the Department for Environment Food and Rural Affairs (DEFRA) and the Scottish government. DASSH manages biodiversity datasets (including images and video) in particular marine benthic survey data.[244] The aims of DASSH are clearly stated on its website as follows:

> DASSH aims to safeguard data (past and future) and make that data available as a national information resource to support marine science and better stewardship of the marine environment. To that end, DASSH provides access to datasets via an on-line catalogue of both metadata and data via this Web site and the National Biodiversity Network (NBN). […] [/] DASSH works with the Marine Environmental Data and Information Network (MEDIN) and collaborates with

---

[240] 'NERC Policy on Licensing and Charging for Environmental Data and Information Products', (p. 2.) *Natural Environmental Research Council (NERC) Website* <http://www.nerc.ac.uk/research/sites/data/policy/nerc-licensing-charging-policy.pdf> [accessed 9 August 2015].

[241] For example of a BODC data licence refer to: 'Academic Licence for the use of data supplied to the BODC: 1 km x 1 km gridded bathymetry for Irish Sea, Celtic Sea and North Channel', *The British Oceanographic Data Centre (BODC) Website* <http://www.bodc.ac.uk/products/external_products/celtic_seas/documents/licence.pdf> [accessed 9 August 2015].

[242] 'Contact details', *The British Oceanographic Data Centre (BODC) Website* <http://www.bodc.ac.uk/about/contact_us/> [accessed 9 August 2015].

[243] 'Online delivery', *The British Oceanographic Data Centre (BODC) Website* <https://www.bodc.ac.uk/data/online_delivery/> [accessed 9 August 2015].

[244] 'Data', *The Archive for Marine Species and Habitats Data (DASSH) Website* <http://www.dassh.ac.uk/data> [accessed 9 August 2015]; [$E_1$].

> existing marine Data Archive Centres (mDACs) to develop and comply with national metadata and data standards.[245]

The DASSH website offers two catalogues for research users one on the DASSH website and a link to the MEDIN Data Discovery Portal.[246]

DASSH provides its bespoke: Terms and Conditions of Data Access and Use on the DASSH website.[247] Briefly, research users are permitted to re-use data held by DASSH for 'use in not-for-profit decision making, research, education and other public-benefit purposes'.[248]

### 4.1.4 UK Hydrographic Office (UKHO): overview

The UK Hydrographic Office (UKHO) was founded in 1795 and operates as a trading fund agency of the Ministry of Defence – also supporting the Maritime and Coastguard Agency.[249] The main aims of the UKHO are clearly stated on the UK government website:

> The UK Hydrographic Office (UKHO) produces nautical publications and services for the Royal Navy and merchant shipping, to protect lives at sea. [/] UKHO is an executive agency, sponsored by the Ministry of Defence.[250]

In the same ways as the other accredited data archive centres, not all their data holdings are part of MEDIN. In the case of the Hydrographic Office, a thematic data archive centre provides its bathymetric survey data.[251] Through the UK Hydrographic Office website, research users can access the UKHO INSPIRE Portal & Bathymetry DAC.[252] These bathymetric survey data include measurements of elevation and depth of the

---

[245] *The Archive for Marine Species and Habitats Data (DASSH) Website* <http://www.dassh.ac.uk/> [accessed 9 August 2015].

[246] 'Search DASSH catalogues', *The Archive for Marine Species and Habitats Data (DASSH) Website* <http://www.dassh.ac.uk/search-catalogues> [accessed 9 August 2015].

[247] 'DASSH Data Policy', *The Archive for Marine Species and Habitats Data (DASSH) Website* <http://www.dassh.ac.uk/data/data-policy> [accessed 9 August 2015].

[248] 'DASSH Data Policy'.

[249] 'About us', *UK Government Website: United Kingdom Hydrographic Office (UKHO)* <https://www.gov.uk/government/organisations/uk-hydrographic-office/about> [accessed 9 August 2015].

[250] 'UKHO: Homepage', *UK Government Website: 'United Kingdom Hydrographic Office (UKHO)* <https://www.gov.uk/government/organisations/uk-hydrographic-office> [accessed 9 August 2015].

[251] 'UK Hydrographic Office INSPIRE Portal and MEDIN Bathymetry Data Archive Centre', *United Kingdom Hydrographic Office (UKHO) Website* <http://www.ukho.gov.uk/inspire/Pages/home.aspx> [accessed 9 August 2015].

[252] 'UK Hydrographic Office INSPIRE Portal and MEDIN Bathymetry DAC' <http://aws2.caris.com/ukho/mapViewer/map.action> [accessed 9 August 2015].

seabed around the UK coast, and are provided as a bathymetric mosaic plotted on a searchable Open Street Map.[253]

In contrast with the BODC and DASSH, all the data held by the Hydrographic Office are public sector data subject to crown copyright.[254] Therefore, the UKHO has a delegation of authority from Her Majesty's Stationery Office to manage and license these data on behalf of the Crown.[255] While the UKHO offers a number of different licences for its data, all the bathymetric survey data are subject to the Open Government Licence.[256] Briefly, research users are permitted to:

- copy, publish, distribute and transmit the Information;
- adapt the Information;
- exploit the Information commercially and non-commercially for example, by combining it with other Information, or by including it in your own product or application.

**You must, where you do any of the above:**

- acknowledge the source of the Information by including any attribution statement specified by the Information Provider(s) and, where possible, provide a link to this licence […][257] [Bullet points and bold emphasis are within the original licence webpage.]

Therefore, while all the bathymetric data held by the UKHO are openly accessible through MEDIN, this is made possible through a different licensing mechanism to the other data archive centres.

At first glance, the UKHO does not appear to fit within the contextual basis of this thesis, as it is not within the UK higher education sector. However, access to bathymetric surveys data is of significant importance to research users within academia. In addition, the inclusion of this data archive centre highlights the complexity of MEDIN's role as a guarantor of data from multiple sources, sectors and organisations across the UK. Moreover, it showcases the complex nature of data re-usage by research

---

[253] 'UK Hydrographic Office INSPIRE Portal and MEDIN Bathymetry DAC' <http://aws2.caris.com/ukho/mapViewer/map.action> [accessed 9 August 2015].

[254] 'Introduction to Copyright Licensing', *United Kingdom Hydrographic Office (UKHO) Website* <http://www.ukho.gov.uk/copyright/> [accessed 9 August 2015].

[255] Copyright, Designs and Patents Act (CDPA) 1988, section 163, *UK Government Legislation Website* <http://www.legislation.gov.uk/ukpga/1988/48/section/163> [accessed 9 August 2015].

[256] 'UK Hydrographic Office INSPIRE Portal and MEDIN Bathymetry Data Archive Centre', *United Kingdom Hydrographic Office (UKHO) Website* <http://www.ukho.gov.uk/inspire/Pages/home.aspx> [accessed 9 August 2015].

[257] 'Open Government Licence for public sector information 3.0', *The National Archives Website* <https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/> [accessed 9 August 2015].

users within academia who not only rely on reliable data gathered by other higher education institutions, but additional authoritative individuals and organisations such as the UKHO.

### 4.1.5 MEDIN's legal framework

By brief examination of the data archive centres' websites (and from the BODC, DASH and UKHO overviews), it is apparent that a number of different licensing regimes are in operation across MEDIN, including: Creative Commons' licences (AGS), the Open Government Licence (BGS, FishDac, the Met Office and UKHO), NERC data licence agreements (BODC), and terms and conditions specified by each data archive centre (AGS, DASSH and the Met Office).[258]

As is to be expected with a portal designed as the UK government mechanism for marine environmental data re-usage, searching for the term: open government licence (on 5 July 2015) via the MEDIN portal returns 4685 data records from the total 9717 records available. The majority of academic research data is likely to fall under NERC data licence agreements or other agreements, rather than the Open Government Licence designed for public sector information. Since more than half of MEDIN's current total records are covered by different licensing regimes, this leads to the first grey area requiring further clarification by the interview participants: to what extent is managing a diverse number of licences problematic? Further to this: to what extent can the MEDIN core team impose specific licensing regimes on its data owners?[259]

---

[258] 'Open Government Licence for public sector information 3.0', *The National Archives Website* <https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/> [accessed 9 August 2015]; 'ADS Terms and Conditions (September 2014)', *The Archaeology Data Service (ADS) Website* <http://archaeologydataservice.ac.uk/advice/termsOfUseAndAccess> [accessed 9 August 2015]; 'MEDIN data submission guidelines', *British Geological Survey (BGS) Website* <http://www.bgs.ac.uk/services/ngdc/management/marine/MEDINDataSubmissionGuidelines.html> [accessed 9 August 2015]; 'Terms and Conditions of Data Access and Use for DASSH', *Data Archive of Marine Species and Habitats' (DASSH) Website* <http://www.dassh.ac.uk/tandc.html> [correct on and last accessed 1 November 2014]; 'Legal', *The Met Office Website* <http://www.metoffice.gov.uk/about-us/legal> [accessed 9 August 2015]; 'Access to Information', '*Centre for Environment, Fisheries & Aquaculture Science' (CEFAS): the 'Department for Environment, Food and Rural Affairs' (DEFRA) Website* <http://www.cefas.defra.gov.uk/publications-and-data/access-to-information.aspx> [correct on and last accessed 1 November 2014].

[259] For more information on MEDIN's data policy refer to: Neil Pittam, Stephen Saxby and Chris Hill, 'Approaches to data policy in the marine sector', *Marine Environmental Data and Information Network (MEDIN) Final Project Report*, Version 1.1 (December 2010)

### 4.1.6 MEDIN's technological framework

The discovery metadata framework has been developed by MEDIN and subsumes other national and international standards of best practice for metadata.[260] These discovery metadata are therefore compliant with the Infrastructure for Spatial Information in the European Community (INSPIRE) Directive, UK GEMINI (GEo-spatial Metadata INteroperability Initiative) 2 standards and ISO19115 – Geographic information – Metadata.[261] MEDIN's technological framework is therefore in part driven by European legislation.

The importance of trans-national sharing of geospatial information, which encompasses some marine environmental data, across Europe led to the INSPIRE Directive. On 15 May 2007, the INSPIRE directive came into force and aims to facilitate greater spatial data sharing between European member states through mandating a shared framework and common metadata standards.[262] Full implementation of the INSPIRE Directive is required by 2019.[263] The obligations of public bodies under the INSPIRE Directive are explained on the MEDIN website:

> Public authorities holding data covered by the directive will have to share their data with other public authorities (e.g. EU institutions). They must allow the

---

<http://www.oceannet.org/library/work_stream_documents/documents/medin_data_policy_study_rep_final_v1_1.pdf> [accessed 9 August 2015].

[260] There are thirty defined elements (plus multiple sub-elements) within the discovery metadata to describe the data quality of a dataset or series. Fifteen elements are mandatory: Resource title, Resource abstract, Resource type, Unique resource identifier, Keywords, Geographical bounding box, Spatial reference system, Temporal reference, Limitations on public access, Conditions applying for access and use, Responsible party, Metadata date, Metadata standard name, Metadata standard version, and Metadata language. Seven elements are conditional: Resource locator, Resource language, Lineage, Topic category, Frequency of update, Spatial resolution, and Conformity. Eight elements are optional: Alternative resource title, Extent, Vertical extent information, Additional information source, Data format, Coupled Resource, Spatial Data Service Type, and Parent ID. For further information see: Becky Seeley and others, 'Guidance notes for the production of discovery metadata for the Marine Environmental Data and Information Network (MEDIN)', *MEDIN Report*, version 2.3.8 (2014)

<http://www.oceannet.org/marine_data_standards/documents/medin_schema_doc_2_3_8_brief.pdf> [accessed 9 August 2015].

[261] 'MEDIN discovery metadata standard', *The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/marine_data_standards/medin_disc_stnd.html> [accessed 9 August 2015]; 'ISO 19115-1:2014 Geographic Information – Metadata—Part 1: Fundamentals', *International Organization for Standardization (ISO) Website* <http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798> [accessed 9 August 2015].

[262] 'About INSPIRE', *Infrastructure for Spatial Information in the European Community (INSPIRE) Directive Website* <http://inspire.jrc.ec.europa.eu/index.cfm/pageid/48> [accessed 9 August 2015].

[263] 'About INSPIRE', *Infrastructure for Spatial Information in the European Community (INSPIRE) Directive Website.*

> public to view data for free, buy the data for download and use over the Internet, and must comply with technical implementing rules to improve consistency. MEDIN is registered as a Spatial Data Interest Community and provides MEDIN INSPIRE updates on the INSPIRE implementation to members.[264]

UK GEMINI 2 'is a specification for a set of metadata elements for describing geospatial data resources for discovery purposes'.[265] These standards have been developed, managed and released as open standards by the Association for Geographic Information (AGI) to provide greater metadata compatibility between individuals and organisations sharing data. These standards are continually updated; UK GEMINI Version 2.3 is expected to be published during summer 2015.[266]

Research users have a wide choice of discovery metadata formats, as these metadata are available through the MEDIN portal in the following five machine-readable versions: directly displayed on the webpage; in MEDIN format; in MEDIN format – XML version 2.3.; in Dublin Core format (XML); and, in GCMD DIF format – XML version 9.4. As these discovery metadata are expressed in non-propriety, common and widely supported formats, research users do not need to purchase software or learn new technological skills to use these five types of discovery metadata. MEDIN also provides the user community with a dedicated metadata helpdesk.[267]

The interviews are used to probe three key areas not addressed within the primary source materials: (a) the advantages and disadvantages of this discovery metadata framework; (b) the impact of prescribed standards enforced by the INSPIRE directive; and (c) the extent to which the marine community has adopted these standards.

---

[264] 'International data initiatives', *The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/online_data_by_theme/international_data_initiatives/> [accessed 9 August 2015].

[265] 'GEMINI', *Association for Geographical Information (AGI) Website* <http://www.agi.org.uk/uk-gemini/> [accessed 5 July 2015]. Les Rackham and Rob Walker, Metadata Guidelines for Geospatial Data Resources - Part 1: Introduction (Association for Geographical Information, September 2010) <http://www.agi.org.uk/uk-gemini/> [accessed 9 August 2015].

[266] 'GEMINI', *Association for Geographical Information (AGI) Website.*

[267] 'MEDIN discovery metadata standard', *The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/marine_data_standards/medin_disc_stnd.html> [accessed 9 August 2015].

### 4.1.7 MEDIN's socio-cultural framework

MEDIN is open to all those with an interest in marine data and information.[268] The amount of discovery metadata it holds is therefore increasing:

> During 2012-13, the number of portal metadata records describing data sets increased from 2,600 to over 4,200 records. The number of portal users has remained steady at just over 12,500 visits in the year, this is expected to increase again once the portal upgrade is completed and the improved portal publicised.[269]

All thematic data archive centres are subject to twelve minimum standards of best practice prescribed by: the MEDIN Accreditation Process for Data Archiving Centres.[270] A data archive centre cannot operate within MEDIN unless it has been approved by the executive team and continually meets the mandated standards. Accreditation is an integral part of the MEDIN model. The twelve MEDIN accreditation procedures are as follows:

1) Adherence to e-GIF and appropriate international principles
2) Data collection according to defined quality principles and accepted procedures
3) Quality assurance of the collected data
4) Databasing and banking with appropriate metadata standards
5) Auditable process for long term custodianship and updating of data sets, with appropriate disaster planning
6) Making datasets freely available wherever possible (not necessarily at zero cost)
7) Committed to raising awareness of the holdings
8) Committed to promoting the use of the data
9) Committed to advising third party organisations collecting similar types of data on procedures, and providing data-banking (warehousing) and curation facilities for such similar data from other sources
10) Committed to, and focus on, customer service
11) Generally exhibiting evidence of expertise and a track record in the scientific area of the data

---

[268] 'Joining MEDIN', *The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/joining_medin/> [accessed 9 August 2015].

[269] MEDIN Annual Report (2012-13), *'The Marine Environmental Data and Information Network (MEDIN) Website*, p. 2 <http://www.oceannet.org/library/key_documents/documents/medin_annual_report_201213_final.pdf> [accessed 9 August 2015].

[270] 'MEDIN work stream documents – MEDIN DAC Accreditation process', version 1.0, *The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/library/work_stream_documents/> [accessed 9 August 2015].

12) Committed to return of data holdings to originators, or lodging with an alternative and suitable repository, if the DAC becomes unsustainable [271]
[Bullet points added by thesis author.]

These twelve accreditation procedures provide the thematic data archive centres with the initial foundations of best practice for safeguarding diverse marine environmental datasets from multiple originators, contexts and sources within MEDIN.

## 4.2 MEDIN: people selected for interview

**Mr B.** is one of the five member MEDIN core team who is directly involved with the management and co-ordination of MEDIN's activities, such as the development of the MEDIN portal, discovery metadata, and the working groups. [$B_1$] **Mr B.** was selected principally for his data management expertise.

**Ms E.** is a data manager at DASSH who began her career as a marine biologist. She has extensive data management experience, particularly with regards to biodiversity databases, and geographical information systems. [$E_1$] **Ms E.** was selected principally for her data management expertise.

**Mr N.** is a technical member of external relations at the UKHO. As a member of a number of MEDIN working groups, he actively influences data policy and standards at MEDIN. [$N_1$, $N_3$] **Mr N.** was selected principally for his data policy expertise.

**Dr S.** is a physical oceanographer at the National Oceanography Centre, University of Southampton (NOCS). She has past experience as a data manager for an international, oceanographic data reusability programme. NOCS is a world-leading institution for marine science research with connections to MEDIN. During the interview, it was revealed that she does not specifically utilise the MEDIN portal as a research user. However, this is noted as an advantage, as she is the only interview participant outside the MEDIN network. [$S_1$, $S_2$] **Dr S.** was selected principally for her academic research expertise.

**Mrs T.** is an intellectual property and licensing officer at the UKHO. She was selected due to her considerable legal expertise, and ensures the bathymetry data

[271] MEDIN work stream documents – MEDIN DAC Accreditation process', version 1.0, *The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/library/work_stream_documents/> [accessed 9 August 2015].

released through MEDIN is compliant with intellectual property law and licensing. [$T_1$, $T_2$] **Mrs T.** was selected principally for her legal expertise.

**Mr W.** is a data scientist at the BODC with substantial technological knowledge. He helps research users to gain access to marine environmental data held by the BODC. [$W_1$] **Mr W.** was selected principally for his technological expertise.

**Mr B., Ms E., Mr N., Dr S., Mrs T.** and **Mr W.** were interviewed during 2012. The interviews were approved by the University of Southampton Management Ethics Committee and adhered to the planned methodology (see Chapter 3 for ethics notice and other information).

## 4.3 MEDIN: interview materials

### 4.3.1 MEDIN's provenance metadata issues

#### 4.3.1.1 Discovery and signposting

It is clear that MEDIN is built around a discovery metadata framework [$B_9$], which all the interview participants describe positively as essential [$B_9$], vital [$E_{17}$, $N_6$], critical [$E_{17}$], and important [$B_9$, $N_6$, $S_{17}$, $T_{16}$, $W_{17}$].[272]

The interview participants refer to the core function of discovery metadata as signposting [$S_2$, $T_3$, $W_2$]; a term that does not appear within the primary source materials.[273] Signposting is defined as the process of signalling the location of existing marine environmental data to research users through discovery metadata [$T_3$]. As all the marine environmental data are dispersed across multiple organisations, research users would find it time-consuming and difficult to locate data without signposts [$S_2$, $T_3$].

---

[272] **Mr B.** emphasises the importance of discovery metadata:

> It's [provenance metadata] essential – well it's at the core of everything, basically is that *er* you've got to know who's produced the data – so that's one of the requirements for filling out the metadata is – what its source is, and what's been done to the data to *erm* – in the processing. So it's – it's fundamentally important to provide assurance in the quality of the data and to allow the user to – *erm* to know what's been done to the data, so it can be reused. [$B_9$]

[273] The interview participants re-iterate some of the other key advantages of provenance metadata raised in Chapter 2, such as for transparency, for scrutiny, to have confidence to re-use data within a safety critical situation (such as navigation) and to record the data source and (field work) equipment used [$N_6$, $S_{17}$]. **Mr N.** states:

> […] The absence of metadata […] is the difference between being – having full confidence in the data and being able to use it in a safety critical *um* situation, such as we [Hydrographic Office] have with keeping navigational charts up to date. *Um*, or –or not being able to use it because you can't trust it. […] [$N_6$]

Discovery metadata further indicates the connections between related datasets that are dispersed across multiple thematic data archive centres, as **Ms E.** states:

> […] The whole concept of MEDIN, this idea of distributed data archive centres [and] a – a central metadata portal – without metadata it wouldn't hang together. […] You couldn't link a metadata record to a – a dataset, or link several datasets together where they're collected under the same survey. […] some datasets we collect have *er* a sort of chemical component, a biological component, geology component – and with the metadata it allows us to store these within these thematic centres of excellence, but also they can be aggregated through the metadata. […] [$E_{17}$]

By signposting related (yet in some instances distributed) datasets through a single portal, the discovery metadata is able to bring together key datasets. This case study demonstrates that metadata are broader than just recording the origins of a dataset [$T_{16}$]; metadata are required for signposting too.[274]

While it is the responsibility of the researchers to provide provenance metadata when depositing data at a thematic data archive centre [$S_{12}$], it does not need to be compliant with MEDIN discovery metadata standards on submission. The MEDIN primary source materials reveal that the MEDIN core team offer support to data originators through the discovery metadata tool and helpline. However, the interview with **Dr S.** reveals that the thematic data archive centres offer an additional layer of support [$S_{12}$]. **Dr S.** offers an example where physical oceanographers write cruise reports (normally in the form of a Word document) that contain all the provenance metadata [$S_{12}$]. These cruise reports are usually submitted to the BODC who will extract the provenance metadata and enter it into their databases in the correct format on behalf the data originators [$S_{12}$]. This is advantageous, as the provenance metadata are scrutinised by an independent party and recorded by individuals who are accustomed to the discovery metadata standards. However, as will be shown by the subsequent case studies not all researchers have such a high level of provenance metadata support.

By providing a transparent and traceable audit trail, provenance metadata protect researchers from allegations of academic misconduct [$W_8$]. This is a point that was not raised within the primary source materials. **Mr W.** expands:

> […] There's a real concern about how to *er* reinstate somebody's scientific reputation if they've been wrongfully or maliciously accused *erm* of scientific

[274] As the majority of the bathymetry data are obtained through the 'Civil Hydrography Programme [...] it doesn't really matter which contractor did the work, or which survey vessel did the work, because the standards are set by us and the MCA' [$T_{16}$]. **Mrs T.** asserts that for the information about the origins of these data are not as important as the date the survey was done and the co-ordinate [$T_{16}$].

> malpractice. *Erm* so, any – anything we can do to *erm* produce an audit trail in the – in the production of a peer-reviewed scientific paper *erm*, I think would be enormously beneficial. *Erm*, I mean for scientific reasons as well – to *erm* encourage further research *er* and to improve the quality of research. […] [$W_8$]

Provenance metadata are not only a tool for research users and data managers to assess the quality of an academic research dataset, but a record that data originators can refer back to and safeguard their good research practice for the future.

### 4.3.2 MEDIN's legal issues

#### 4.3.2.1 Diverse data licence agreements

MEDIN appears to have a strong legal framework as each dataset falls under a data licence agreement [$B_4$, $E_6$, $W_{15}$], and MEDIN actively meets national and international legislative requirements.[275]

As a result of MEDIN's cognisance of the legal issues pertaining to academic research data re-usage, it was revealed from the interviews that the mandatory utilisation of licensing results in MEDIN operating within the current copyright framework without difficulty.[276] Furthermore, as shown by the primary materials, the majority of academic research data collection is funded by NERC and therefore such

---

[275] **Mr W.** contends that BODC 'don't supply any data *erm* that's not under a data licence agreement', as data originators (particularly commercial data suppliers) are more likely to submit their data when they are re-assured by binding terms and conditions [$W_{15}$]. **Mr B.** states:

> […] MEDIN doesn't own any data itself – so MEDIN is a – a co-operative set-up, if you like. So, the data still belongs to whoever it belonged to in the first place. So, it could be whoever claims the IPR for it […] it could be NERC – the Natural Environmental Research Council […] it could be a government department, or it could be a commercial organisation. So MEDIN doesn't interfere with the ownership of the data, it just helps make the data available. *Erm* and I guess what we do is we try and encourage people [to be] as open as possible with their licensing […] and we encourage people not to limit access to their data. […] [$B_4$]

**Ms E.** states:

> […] It's all third party data that we hold within DASSH. *Um*, the data that comes into DASSH we don't take ownership of, we're custodians of it. *Um* so they [data owners] sign our terms and conditions, *um* and there's differing levels at which *um* they provide data to us. Either we archive it and it doesn't go anywhere, it's – you know it just sits within our systems, or ideally *er* it's made as freely available as possible, particularly as it's collected with public money then there's a – an obligation to do that. [$E_6$]

[276] **Mrs T.** states: 'In fact, by the nature of – of bathymetry data, which is survey data, *er* which is collected electronically, it's very, very doubtful that there is any copyright in that, but there is a database right' [$T_7$]. The EU *sui generis* database right is quite confusing, as it is not well-known, there is very little case law. There is not an international convention, as there is for copyright and patents, and it does not exist in some countries [$T_8$]. From the interviews, it appears that legal factors do not significantly hinder the utility of the MEDIN portal and the data archive centres – '[copyright and licensing] in our day-to-day operations, it's not very restrictive'. **Dr S.** adds that '[legal issues are] not something that I've ever come across – no, never' [$S_{14}$].

researchers have to comply with the NERC data policy, which clearly stipulates that data should be released openly where possible and fall under a data licence agreement [$W_{13}$].

While the legal framework seems to be robust, on examination of the MEDIN primary source materials, a (potential) key problem was flagged for further investigation within the interviews – to what extent is MEDIN able to successfully manage a number of diverse data licence agreements in operation within the thematic data archive centres.[277] This was recognised as a potential issue by **Mr B.** who would like to see the creation of a list that contains 'a more limited number of standardised licences' [$B_{25}$].

Moreover, it appears that the increased use of a single licensing regime for public sector data held by MEDIN – the Open Government Licence – is having a positive impact on data management and re-usage, as **Mrs T.** explains:[278]

> […] The Open Government Licence will have helped, because many of the contributors will be public sector bodies who are very strongly encouraged to use the open government licence. *Um*, but there will be private sector contributors as well of – of data and *um* they will make their own decisions. […] It's a balancing act between commonality, which makes it easier for the user, but if it – if you're mandating things or being too prescriptive then you get less data. […] I think it's likely that there will remain a variety of licences, but I think a significant portion of the data will be open government licence. […] [$T_{18}$]

Due to the apparently successful uptake of the Open Government Licence [$T_{17}$],[279] it would be useful to consider whether an equivalent licence could be developed for the UK higher education sector and disseminated through Research Councils UK for instance. While this is outside the scope of this thesis, it could be a potentially fruitful area for further research.

MEDIN is not a legal organisation or 'a legal entity […] it can only point […] people towards good models' [$B_{26}$] and does not, therefore, have a role in the creation of data licence agreements. For example, **Mr B.** contends that the government is better

---

[277] While all the bathymetry data held by the UKHO are covered by an Open Government Licence, the UKHO operates additional licensing regimes for its other types of data and products, such as: 'a free of charge licence for non-commercial use or low value commercial use' and an online licensing system [$T_{12}$]. These licences 'only last a year […] and if the user wants to continue to use the data they go online and apply for another licence when it expires – the open government licence doesn't expire' [$T_{12}$]. Although these licences existed before Creative Commons, the Hydrographic Office has followed the developments of Creative Commons and the Open Government Licence, and have modified the language of their agreements to make them more familiar [$T_{13}$].

[278] Like many other government bodies, the Hydrographic Office have adopted the use of the Open Government Licence [$T_6$].

[279] **Mrs T.** contends 'the open government licence is very simple and clear, and because it's very widely used *er* many users of the DAC, I think, will already be familiar with the open government licence' [$T_{17}$].

placed to define licences for its own organisations [$B_{26}$]. MEDIN only has limited control over the existing licensing regime, as it appears unable to make its own MEDIN data licence agreement.

While a reduction in the number of different licensing regimes is favourable, minimising options for permissions and access may discourage some data owners from depositing their data at the data archive centres. **Mr B.** further contends that, with more limited choice, data owners may favour more restrictive licences (such as, for non-commercial re-use) even though there is limited justification [$B_{25}$]. Therefore, MEDIN safeguards the interests of its data owners by supporting an assortment of licensing regimes suited to the multiple sources, contexts and types of marine environmental data held by the thematic data archive centres.

Single licensing regimes do not result in the same permissions being granted either and may have to be amended to reflect an organisation's responsibilities. For example, the UKHO had to slightly amend the Open Government Licence.[280]

#### 4.3.2.2 Restrictive licensing

From the MEDIN primary source materials, it is apparent that MEDIN aims to openly release all discovery metadata (regardless of any restrictions on re-usage of a particular dataset) [$W_{20}$], and actively encourages all marine environment data to be openly released where possible [$B_7$, $W_{20}$].[281] From the interviews (in 2012), it is estimated that two thirds of marine environmental data held by the thematic data archive centres are openly released, and the other third are embargoed [$W_{19}$]. For instance, two out of one-

[280] **Mrs T.** states:

> […] [The Hydrographic Office] have had to modify the Open Government Licence very slightly. And that's because our data can – could be used to create navigational products. […] And we have a responsibility to try to ensure that safe navigational products are available and that they can be distinguished from others. *Er* because the government has a liability – a potential liability in case of *um* a passenger vessel going down or a [sic] oil tanker *um* spilling millions of tonnes of oil onto beaches and things like that. So we've had to modify the open government licence to restrict that users can't use the data to create something which is pretending to be an official navigational product. [$T_6$]

[281] It is argued that marine environmental data collected by taxpayers' money data should be available to the public [$B_7$]. **Mr B.** expands on this:

> […] We could encourage everybody to make raw data as freely available as possible because, by in large, once they've got the use out of it by writing *um* a PhD, or writing a scientific report, or deciding where they're going to put their wind farm or their oil well – *um* once that primary use is out of the way, then I think the raw data have – tend to have very little commercial or – value. […] [$B_7$]

thousand biodiversity datasets at DASSH are embargoed – this appears to be because the datasets contain information about sensitive habitats [$E_7$].

It is evident from the MEDIN interviews that restrictive licensing and embargos have an important role for safeguarding marine environmental data research [$B_8$, $N_9$, $S_7$]. Academics often want to publish and analyse their data before it is made publically available, therefore short term embargos (that usually last for a couple of years) facilitate this reasonable academic endeavour [$B_8$, $N_9$, $S_7$].[282]

Restrictive licensing is not only required for researchers to exhaust the primary usage of a dataset, but to prevent the release of sensitive data. While it is not the main focus of this case study to confront the issues that arise from sensitive data (this is addressed in Chapter 6: FLLOC and SPLLOC Case Study) nor is it a key issue at MEDIN, a very small minority of data held by the thematic data archive centres are restricted due to their sensitive nature. Out of the third of data that are embargoed, approximately one per cent of those data are completely restricted [$B_8$, $B_{12}$, $W_{19}$]. A small proportion of biodiversity data held by DASSH is classed as sensitive data. This is because their disclosure would cause a risk of disturbance and/or destruction to vulnerable marine species and habitats by inquisitive individuals [$B_{12}$].[283] These sensitive raw data are indefinitely embargoed [$B_8$], but modified versions of these sensitive data are still made openly accessible instead, as **Ms E.** illustrates from data redaction at DASSH:

> […] So with *erm* data relating to […] species distributions, quite often it's got counts of species. So you've found ten of this species, ten of this, twenty of this one. What [DASSH] we've done is, with the agreement of the data provider, is change that so it just says: 'this species was found here', but it doesn't give the counts. It just *um* changes it to a presence/absence, 'yes it's here, no it isn't' type dataset. So we've – we've negotiated the – the capacity to release it in a modified form. […] I still think that's better than not releasing it. [$E_7$].

Therefore for the majority of sensitive datasets access to the raw data is restricted, and modified data are released instead [$B_8$, $B_{12}$] in the following key instances: commercially sensitive data [$B_8$]; socio-economic data such as fish catches [$B_8$]; and,

[282] There is concern amongst 'academics that the commercial companies will financially benefit from their academic research' [$E_{19}$]. But their research is 'funded by UK government to do this work for the benefit of the UK […] it's for the wider good' [$E_{19}$].

[283] It must be noted that **Mr W.** has not dealt with any sensitive data issues at the BODC [$W_{22}$]. Chapter 6 centres on and critically analyses issues that arise from making sensitive and personalised data re-usable.

vessel tracking information (details of where a vessel has been), which is the only example of personalised data held by MEDIN [$B_{12}$].

#### 4.3.2.3 European harmonisation

From the interviews, it appears that European politics has significantly driven the standardisation of marine environmental data across Europe [$E_{22}$] through key European legislation – comprising: the Environmental Information Regulations, the Freedom of Information Act, the INSPIRE Directive [$E_6$], Marine Strategy Framework Directive [$E_9$], and the Re-use of Public Sector Information Regulations [$T_{21}$].[284] Given MEDIN was established the year after the INSPIRE Directive came into force in 2007, it has been well-placed to respond by directly implementing this change and raising awareness across the community.

As there are no political boundaries in the marine environment ('the fact that a fish is over in Belgium doesn't stop it swimming over into German waters' [$E_{22}$]), legalisation sought to tackle duplicate data collection and co-ordinate data sharing across Europe, as 'each member state in Europe was collecting data in its own ways and its own formats' [$E_{22}$].[285] The INSPIRE Directive appears to be at the forefront of this standardisation [$E_4$], and enables DASSH to interact with European partners and enter into collaborative projects [$E_3$].[286] The INSPIRE Directive has also had a positive impact on the Hydrographic Office, as **Mr N.** states:

> […] Our metadata that we [the Hydrographic Office] used to hold was not necessarily INSPIRE compliant because we were holding it for purposes of creating navigational products not for general external use. So there's been quite a bit of work we've had to do to bring our metadata into a state […] which is

[284] According to **Ms E.**, three areas of law significantly impact on DASSH: (1) the Freedom of Information Act; (2) the Environmental Information Regulations; and, (3) the INSPIRE Directive [$E_6$]. The Environmental Information Regulations place an obligation on the Marine Biological Association to provide access to data where it meets the requirement of the public interest test [$E_{19}$].

[285] For example, before the Marine Strategy Framework Directive, there was no single repository, standard format and standardised way of collecting marine litter and underwater noise data within Europe [$E_9$].

[286] **Mr N.** adds to this:

> […] It is however also acknowledged that there is data which we [Hydrographic Office] hold, *um* which is of interest more widely than just for navigators. […] Particularly with the advent of INSPIRE regulations, *er* we've been focusing on which datasets we have a responsibility for that we should be making available as part of our obligations under INSPIRE. *Er* so we've looked at where we hold data, which is not held by other people, or at least where we have the best copy of that kind of dataset, or where we the produce data ourselves. I think it's very important to make clear that we don't produce data ourselves by in large. We receive data from other organisations and […] we process it, we examine it, we compare it to data that already exists, in order to keep our navigational products up to date. […] [$N_3$]

> compliant with the INSPIRE directive. *Er* so it's had quite a significant effect. In terms of making data available, we – we already did anyway. *Er* we are subject to the Re-use of Public Sector Information Regulations […] which already effectively required us to make the data available for re-use *um* for virtually everything that we hold. And, that's been in force since 2005. […] The main impact of INSPIRE has been on metadata. [$T_{21}$]

The interviews further reveal that INSPIRE does not capture all marine environmental datasets, and there is a degree of uncertainty over which datasets fall under its scope [$N_{20}$,$T_{21}$]. For instance, the UKHO was unclear about its bathymetric survey data as within the scope of INSPIRE, but sought to make these datasets compliant [$N_3$, $N_4$].

The MEDIN portal website makes clear that the discovery metadata is compliant with INSPIRE standards. The GEMINI or INSPIRE standards often changed every six months, but these standards are now maturing [$B_{19}$]. MEDIN has a vital role as 'a fantastic tool and resource' [$E_4$] to demonstrate the ways in which individuals and organisations are required to make their data available in a certain structure that meets institutional, contractual and other legal obligations.

While European legislation is having a seemingly positive impact on data sharing within the marine environmental community at national and international levels, it is unclear to what extent such legislation would be an option to change existing data sharing practices in other disciplinary domains, where significant political drive to standardise data re-usage is currently lacking.

#### 4.3.2.4 Research misconduct

While there are no known examples of misuse of data within MEDIN [$B_6$, $T_6$], DASSH [$E_{15}$] or generally [$N_5$, $S_{14}$], some academic researchers in the marine sciences are still concerned that their data may be misused despite prescriptive terms and conditions within data licence agreements [$E_{16}$]. MEDIN is however unable to ensure that terms of data licence agreements are adhered to directly [$W_{11}$]. This is because it relies on its user community to report back their suspicions where an individual and/or organisation appears to have breached the terms of the data licence agreement. Hypothetically, where misconduct is proved upon investigation by MEDIN, it would be able to 'refuse a customer [research user] any further data' [$W_{11}$]. Therefore, misuse of data at MEDIN is self-monitored.

MEDIN makes marine environmental data openly accessible on the basis of goodwill and trust [$W_{11}$], without requiring any form of research user registration, which can be quite burdensome as **Mrs T.** suggests:

> […] We [Hydrographic Office] did at one stage consider requiring users to register. *Um*, and the main purpose for that originally was so that we would be able to potentially police misuse of data. *Er*, having gone to the Open Government Licence where there really isn't the concept of misuse of data, we – we've dropped the idea of registration. Registration carries with it significant *um* rules and regulations, because of *um* data protection and privacy – *um* which in itself results – would result in a significant additional cost *um* in the – for the infrastructure and setting up the DAC. […] Data protection is always something we need to be aware of, and we've side-stepped it by not collecting any personal data at all. […] [$T_{20}$]

In the experience of **Ms E.**, there is a reliance on gentlemen's agreements and good will to resolve such cases of alleged misconduct rather than bringing a legal action against the individual(s) who breach contractual terms [$E_{15}$]. The BODC has never tried to pursue a legal route but, in the very distant past (before the establishment of MEDIN) has refused data to customers [$W_{12}$]. Funders, colleagues, publishers and future collaborators would be wary of working with a known misuser [$E_{15}$]. In theory, therefore, a breach of contractual terms may be strong grounds to bring an action in court, but in practice this is avoided as any litigation is potentially lengthy and costly. A tarnished academic reputation through misuse of academic research data may be more of a disincentive for research users than legal proceedings. More positively MEDIN has a transparent policy of its users scrutinising data and reporting back anything irregular. This means that genuine errors and omissions can be highlighted, ultimately raising the quality of data. However, this appears to undermine the legal strength of the data licence agreement. Data licence agreements are disincentives to misconduct, but their lack of enforcement may ultimately undermine this function.

Aside from potential misuse from research users, **Dr S.** raises another potential area for misconduct by data originators – 'NERC […] can, you know, take sanctions against a PI [principal investigator] if they don't submit their data' to a data archive centre [$S_{14}$]. **Dr S.** has no examples of this [$S_{14}$], but as the open data movement becomes more influential (as shown in the literature review) it will be interesting to observe whether this occurs in practice.

### 4.3.3 MEDIN's technological issues

#### 4.3.3.1 Search functionality and web delivery

One of MEDIN's key successes lies in its open, community-led approach to creating its IT infrastructure and its commitment to simplicity. MEDIN is actively meeting the requirements of research users, which are emphasised by **Dr S.**:

> […] I think the critical thing is to make the information and the data format itself as simple as possible. So that people *um* don't have to learn new software […] or buy new software in order to access datasets. I think the – the data files themselves need to contain as much information as possible in a way that's easy for people to just kind of click and open. […] [$S_3$]

From consideration of MEDIN's primary source materials, it appears that MEDIN faces only very minor technological issues, which is subsequently confirmed during the semi-structured interviews [$E_{12}$].[287] However, the interview participants highlight two important areas for further technological development: increasing portal search functionality; and, expanding web delivery of data.

From the interviews, it is evident that the portal search functionality requires further fine-tuning [$W_3$] and there is scope to facilitate more enhanced queries. For instance, while research users are currently able to search for data within a specified region, they cannot find answers to questions such as 'where can I exploit renewable energy more efficiently?' [$W_3$]. Moreover, research users are not yet able to limit their search on the MEDIN portal to unrestricted data only [$W_{18}$].[288]

Although MEDIN is aware of the minor improvements required (at the time of the interviews in 2012) there are no specific plans to confront these shortcomings due to a current lack of resources [$W_5$]. As is the case with all data re-usage, improvements in technology will continue to impact on the future search functionality of the portal [$W_{23}$].[289] It is a possibility that the solution to greater and enhanced data querying may come from linked data [$E_{23}$].[290]

---

[287] **Ms E.** confirms that DASSH have not faced any software or systems problems [$E_{12}$].
[288] Therefore, often data are not available for instant download, and the data contact point at BODC will have to email and negotiate access, which is quite laborious [$W_{18}$].
[289] In the near future, therefore, **Mr W.** would like to see 'massively more data available', and software development for enhanced data queries [$W_{23}$]. In the future, general developments in IT will be crucial also, such as the cost of greater storage; increased speed; the quicker conservation of data; and, more powerful searching capabilities [$W_{23}$]. Furthermore, data collection (field work) might be changed by a rise in portable mobile devices that would feed directly into the data archive centres [$E_{24}$]. Politicians are also increasingly interested in climate prediction models [$S_{24}$].
[290] DASSH are aware of linked data [$E_{24}$].

As they had to do before the Web was a worldwide phenomenon, research users still have to phone the data owner/manager to arrange for physical delivery of data via a disk, portable hard-drive or other medium [$B_{17}$, $E_{13}$] for very large volumes of data [$E_{13}$]. While smaller datasets are often available for instant download from the data archive centre's website [$E_{13}$], it appears that web delivery is a definite area for future improvement, as **Mr B.** suggests:

> […] There's quite a lot you can get through the Web – so you can […] click through and just order a dataset – but […] it's nothing like an Amazon store at the moment. […] That's where a lot of people would like to see us [MEDIN] ending up – is that it's more or less a one-stop shop. You go through portal. You find the datasets you want. You look at the metadata. You check – you kind of reduce that initial set to what you really want. And, then basically you just fill up your trolley with those datasets and get it delivered. [$B_{17}$]

It is anticipated that in ten years' time the majority of data held by the thematic data archive centres will be available for instant download complete with data snapshots for research users to initially preview [$B_{18}$]. However, in the longer term this may involve re-engineering the technological structures underpinning the data archive centres [$B_{18}$]. Third party cloud services do not currently appear to be an option (as data storage is out of control of the thematic data archive centres) without a robust cloud service agreement in place. Any third party cloud service must offer high levels of security and data protection; especially where data are stored as closed records and/or are of a sensitive nature.

### 4.3.4 MEDIN's socio-cultural issues

#### 4.3.4.1 Data sharing

While it is clear from the MEDIN primary source materials that data archive centres predate both MEDIN and the Web, and therefore are not new [$S_8$], marine environmental data have not always been made widely available [$S_6$, $S_9$].[291] From the interviews, it is further apparent that there is a considerable historical backlog of important, yet inaccessible, marine environmental data that unless archived constitute

[291] **Dr S.** contends that physical oceanography academics have always been conscientious and deposit their data at national archive centres – they need to know the data is safe for future research, and its availability is not affected whether a researcher retires or changes employment [$S_6$]. **Dr S.** further states:

> […] [Marine environmental data] hasn't been, you know, widely available. So the – the sort of advent of portals and being able to download things from the Web is just brilliant. […] It makes the data much more available to many, many people; which is absolutely the right thing. [$S_9$]

the risk of data loss [$E_{29}$]. However, the digitisation of such data is extremely costly in resources, staff time and finances [$E_{29}$].

Indeed, a significant number of marine environmental datasets have been lost, because data archives did not have the funding to offer a secure archive and conduct quality checks [$W_9$]. Even where data had funding set aside for archival costs – chiefly core datasets collected during specific NERC funded projects – a considerable amount of outlying data gathered during these projects remains at risk from data loss unless the principal investigator actively organised secure archival for these datasets too [$W_9$].

The amount of data is increasing, but the associated costs of data storage are falling; the risk of losing important datasets has also decreased as data archive centres are able to manage more data [$S_8$].[292] For instance, the BODC provides a free secure archive service for data collected during non-NERC funded projects, offering basic quality checks and standard metadata to aid data discovery [$W_9$].

Attitudes towards data sharing within the marine environmental research community have also taken a more positive stance.[293] From the interviews, it is clear that the ways in which marine environmental data are now collected, disseminated and re-used have also changed attitudes towards data sharing in the marine community [$S_{20}$, $S_{23}$, $S_{26}$]. [294] **Dr S.** explains how physical oceanographers require a wider range of existing data to re-use within research projects, and to support the collection of new datasets:

> […] In physical oceanography, you know, the traditional – the old fashioned route was: you go out to sea, you collect a dataset, and then you spend five years analysing it, and write a couple of papers. But nowadays, I think that people expect to be able to use data from the region around them, *um* different – lots of

---

[292] Around fifteen years ago, research users would write or phone the data centres to get access to marine environmental data [$S_8$].

[293] Although it is outside this thesis, the UK government's open data agenda – such as the creation of open.gov.uk and the open government licence – impacts on government bodies to make their data available via MEDIN, which was once viewed as a commercial resource [$B_{16}$]. **Mr B.** explains:

> […] They've [opendata.gov.uk and open government licences] forced *um* MEDIN partners who are government bodies and agencies to – to make their data available. *Um*, because not very long ago, a number of agencies were thinking of their data as a potential – *er* commercial resource. […] If they were set up as kind of semi-commercial bodies […] they said: 'well, we – we have to make some income of all of our resources that's part of our charter', so they – they argued that they actually couldn't make their data freely available […] [$B_{16}$].

[294] **Dr S.** states:

> […] I think even as academics, we make wider use of our data now than we used to, because we will still write the journal articles, *um* and write a data report and that sort of thing. But, I think that we are spending much more time contributing to status reports and impact reports, and providing more sort of analysis, the – the raw data, but also the *um* – the sort of impact analysis of the results that we've collected. […] [$S_{23}$]

> different kinds of data not just data that you've collected from ships. […] I think it's sort of a change in approach to how you do your analysis; it tends to be a bit more wide-ranging then just focusing on one specific, small part of the ocean. […] [$S_{20}$]

As is shown by the MEDIN primary source materials, the majority of marine environmental research is funded through NERC or universities, therefore researchers are subject to the NERC data policy and (if available) the data policy of their higher education institution(s) [$B_5$,$W_{11}$].[295] Therefore, these minimum mandatory requirements have had a positive influence on encouraging increased data sharing. This policy ensures that the widest possible collection of quality data from non-NERC funded researchers is encouraged as part of important data re-discovery in future projects.

While generally the marine community appears to be more supportive of data sharing, from the interviews the culture of sharing within the marine sciences still remains variable [$W_{32}$].[296] Although, many researchers have experienced 'big, NERC funded, multidisciplinary projects' [$W_{32}$], there are a minority who may not want to share, perhaps because they are externally funded or part of specialist research projects [$E_4$, $W_{32}$].[297] Some researchers still need to be shown the benefits of data sharing [$S_{19}$, $W_{32}$]. [298] Therefore, while data policies are able to influence the data sharing practices of those researchers that fall under their remit, they are unable to offer an all-inclusive solution for encouraging increased data sharing across the entire marine community (such is the case within all other disciplinary communities).

It is evident therefore that future re-use appears to have become the norm in the collection of marine environmental data, as many researchers now anticipate that other

---

[295] **Mr W.** re-iterates that BODC is 'governed by the NERC data policy so *erm* in practice we operate to make everything available *erm* as much as possible with – within the constraints of the NERC data policy' [$W_7$].

[296] **Dr S.** is unsure whether the positive sharing culture found within physical oceanography extends across other disciplines in the marine sciences [$S_{11}$].

[297] According to **Ms E.**, unlike bioinformatics and genetics academics that have to work collaboratively and share data, biodiversity academics often work individually or within a small team creating long-term, time-series data at particular sites [$E_4$]. Within the marine biodiversity sector, therefore, there has often been difficulty engaging academics as researchers see it as 'their data and that it's not something to be shared with the community' [$E_4$]. According to **Dr S.**, in the 1990s two key international data sharing programmes occurred within physical oceanography – the World Circulation Experiment and the ARGO Float programme – which highlighted the benefits of making data available for other to re-use, and encouraged future data sharing [$S_{19}$].

[298] **Dr S.** states: 'I don't think you can automatically assume that everybody wants to share their data straight away, you'll have to show people why it's a good idea' [$S_{19}$]. **Mr W.** re-iterates this point by suggesting that academic researchers who experience the benefits of accessing others' definitive version of data are more likely to make their own data re-usable [$W_{32}$].

research users will want to re-use their data.[299] Moreover, it is common knowledge that there is heightened significance placed on the future management of the marine environment from researchers and policy makers, such as concerns over climate change [$E_{25}$].[300]

While academic research is independent of government [$S_{26}$], it is understood that in order to decide on the future management of the marine environment, policy makers require the best available data. As academic researchers grow more cognisant of this wider audience, data are increasingly presented in ways which non-experts can understand [$S_{26}$].[301]

Although MEDIN is the UK government's mechanism for marine environmental data re-usage [$T_6$], MEDIN is not driven by conservationist or exploitation agendas [$B_{21}$, $E_{19}$, $N_{14}$, $S_{25}$] and is not directly involved with the management of the marine environment as that is the responsibility of policy makers [$B_{21}$, $W_{27}$]. MEDIN only aims to provide the most reliable data for future re-usage for whoever requires it, which includes but is not limited to policy decisions taken by UK government bodies [$E_{23}$, $B_{21}$, $N_{14}$].[302]

---

[299] **Dr S.** states:

> […] I think any portal like MEDIN certainly does have an impact. *Um*, because it means your research, as an academic […] can *um* sort of be wider than just the data you are able to collect. […] So if you're doing a project on North Atlantic currents, you're not just going to be using just the data you've been able to collect; you have access to other people's data as well. […] From funding bodies' perspective I think what that means is that *um* you – every dataset becomes – is more valuable really, because it has potentially more use. [$S_{10}$]

[300] **Ms E.** states:

> […] As increased pressures on the marine environment with *um* things like climate change and rising sea level *um* those are going to be issues that people need more information. *Um* so it's really just a question of whether the existing structures can scale-up to deliver the increased amount of data that's going to be available. [$E_{25}$]

[301] **Dr S.** states:

> […] Particularly, over the last ten years, we've [academics within physical oceanography] made more of an effort to summarise […] our data and our findings in a way that government *um* will understand. […] So that we're contributing to policy development. So we're not setting policies, but we're trying to provide the information that the people who set policies will understand. […] [$S_{26}$]

[302] **Ms E.** expands on this:

> […] [The Marine Biological Association that hosts DASSH] just provide – provide the evidence and our professional interpretation of that evidence. If then, that goes off to be used for *er* I don't know, setting up a marine conservation zone, or for mandating a drilling platform – as long as we can defend the information that we've given out *um* that's sort of where we […] draw the line. […] Again it's – it's a concern, scientists don't want to be seen to be supporting potentially damaging activities with their data, because it might be that their data […] is incomplete, *er* hasn't been *um* as thoroughly […] quality assured as they would like. So, in those cases, we would *um* release it with a *um* sort of caveat or a statement explaining the status of that data at that time. [$E_{19}$].

#### 4.3.4.2 Gather data once and use many times

From the primary source materials and the interviews, it is clear that MEDIN was formulated to address a number of key data sharing issues within the marine community, such as a lack of standardisation, limited accessibility, incompatible formats, and deficiency of provenance metadata [$B_2$, $B_3$, $E_2$]. Moreover, the interviews further reveal that the utilisation of particular commercial software packages and data loggers also precluded the wider re-usage of marine environmental data.[303]

Marine environmental data are (generally) expensive to collect, as they often require specialised equipment, fieldwork and vessel hire. Inadequate data sharing across the marine community causes data duplication, a waste of resources and unnecessary costs [$E_2$, $N_2$].[304] Therefore, the principle: gather data once, use many times, was an important point raised during the interviews; this principle lies at the heart of the MEDIN discovery portal and its network of accredited thematic data archive centres [$E_2$, $N_2$].

MEDIN not only counteracts the needless collection of duplicate datasets [$B_2$, $E_4$, $W_6$], but further prevents organisations and individuals duplicating efforts to share and manage data [$E_4$]. By offering and encouraging data collectors/owners to deposit their data within a single, authoritative mechanism, MEDIN is able to reduce the

---

[303] As part of **Dr S.**'s past experience as a data manger with an oceanographic research team in the 1990s and early 2000s, this team surveyed research users to find out their preferred data formats. Many survey respondents wanted the simple data format ASCII where no additional software was required [$S_3$]. Therefore, the oceanographic research team provided their data in ASCII and another easy to use format [$S_3$]. **Dr S.** described asking research users about formats as 'really important' [$S_3$]. **Mr B.** further states:

> […] What we're [MEDIN] trying to do is develop common approaches to basic data processing – so that they're [marine data] available again without having to – to buy commercial software. *Um* – and the whole point of this data archive centre framework is that the data are presented in common formats. So they can take – the data archive centres can take data in whole range of different formats – but they make it easier for other users to use the data – because they present them, they serve them up in common – *er* commonly used, different formats – and not in kind of bespoke, or industry specific formats. [$B_3$]

[304] **Mr N.** explains the principle of 'gather data once, use many times':

> […] Gather data once, use many times. I think there was an awareness *er* that data was being gathered by various organisations and only used by them. And so, you could have *um* the possibility of expensive surveys being carried out by an organisation in the same area *er* either at the same time or very soon after that another survey had already been carried out, so that was seen as very wasteful. [$N_2$].

**Ms E.** re-iterates the principle of 'gather data once, use many times' :

> […] It's about *er* trying to make datasets comparable so that people aren't spending *er* a lot of money re – re-collecting data that's already been *er* collected. I mean the – the cost of marine data collection is massive. […] I think the mantra of MEDIN is use – is 'collect once, use many times'. *Um* that obviously makes it more cost effective, and [the UK] government's very keen if they're paying for *um* work to be done that it has a long-term use and – and applicability, rather than just being sort of pigeon-holed for a single project. [$E_2$]

number of places data are held [$B_2$] whilst increasing the discoverability and improving the ways in which data are managed.[305]

In theory many individuals and organisations are now willing to share their marine environmental data, but some of their existing organisational practices do not support present standards of best practice [$B_{15}$]. For individuals aspiring to change these incompatible practices, it is extremely challenging to justify the extra costs of altering formats, metadata standards and the publication process to their organisation via a strong business case [$B_{15}$].[306] For instance, some of the MEDIN thematic data archives – such as the UKHO that was founded in 1795 – have been custodians of marine environmental data for many years, whereas DASSH (founded in 2005) has been able to adapt to new changes more quickly [$E_{10}$, $E_{26}$].[307]

Through its diverse thematic data archive centres, MEDIN has indirectly inherited a diverse, and historic, collection of data and data management practices [$E_{26}$]. While it would (potentially) be easier to manage a centralised database of marine environmental data configured by current standards, MEDIN has to work within existing organisational structures [$E_{26}$] and their longer histories. The breadth of expertise offered by diverse thematic data centres would be difficult to replicate in a single, centralised database [$E_{26}$].

While data policies, national and European standardisation (such as the INSPIRE directive) have largely helped MEDIN to achieve the principle of gather data once, use many times, the research in this case study has also identified community involvement as key to MEDIN's overall success [$B_{15}$**,** $B_{22}$**,** $E_{10}$, $E_{27}$].[308] MEDIN is built

---

[305] Through community agreed and adopted standards of best practice, the core team and thematic data archive centres support individuals and organisations to transform their current data into the most re-usable data format possible [$B_3$]. MEDIN and its thematic data archive centres are important as they safeguard data long after projects finish and their websites disappear [$E_{29}$].

[306] **Mr B.** further states:

> […] It's actually more about – *um* changing the way you manage data internally. So, changing your internal practices so that they match better what's going on elsewhere. […] That's why it's quite useful when there's a bit of a top-down push from government to make – to adopt open government licences for instance – and […] people publish their data. [$B_{15}$].

[307] **Ms E.** adds:

> […] [The MEDIN core team] depend on the MEDIN community, it's – it's a partnership. So *um* the – each data archive centre is involved […] in a working group […] so each of these different strands of MEDIN – this is really steered by the – the community. *Um* it has to be that way. Otherwise, if you just have a small group trying to impose something it would […] never have got off the ground. […] Because there's – there's so many established data archive centres […] that have very entrenched ways of doing things that that there's a certain amount of sort *um* sort of squeezing and juggling to get these things to fit. […] [$E_{10}$]

[308] **Mr B.** states:

on the simple principle of getting 'everyone in the same room' to discuss, influence and adopt common standards of best practice [$B_{22}$]. **Ms E.** stresses the importance of community involvement in the following interview extract:

> […] You need to get all the – all the relevant users and holders of the data that you're dealing with *um* round the table. [...] Try to build that consensus is – is [an] incredibly difficult task and one that MEDIN have done very well. […] There's a danger in these sort of partnerships that people sort of say: 'yeah, yeah, it's brilliant, it's great', but don't actually […] commit to it. But I think that's something that comes with time, because eventually you build a critical mass […] of organisations, and data, and resources – that if you're not involved in it you're actually losing out. […] [$E_{27}$]

From the MEDIN primary source materials, it is evident that MEDIN involves a large number of stakeholders and partner organisations. The multiple players in the marine environmental community all have a significant influence on the remit, shape and capabilities of MEDIN and its thematic data archive centres [$B_{22}$]. Through this extensive community involvement, the standards developed for marine environmental data have become national standards which are implanted not only within the marine environmental community, but UK government [$E_{27}$].

#### 4.3.4.3 Awareness and portal multiplicity

While MEDIN is based on a strong community approach and has many sponsors and partner organisations, it appears from the interviews that the MEDIN portal is not sufficiently well-known within the physical oceanography discipline [$S_{29}$] and is underused [$W_{25}$].[309] **Mr W.** states: 'I think more academic researchers need to know *er*

---

> […] It's just the importance of getting everybody working together *um* … so that they – internally they recognise the – the importance of what's being done, rather than trying to sort of force a top down solution on it. […] We [MEDIN] developed the structure we did because it kind of suited the environment. So, we have a working group specifically on standards. So, you know, those people in – in the marine field who are interested in standards can get involved in that. […] We don't have kind of one or two experts that tell everybody else what to do. We – we get the organisations to actively participate in the working groups, in the different areas, in standards and in developing the portal. […] [$B_{22}$]

Without a government mandate many organisations may not have been involved with MEDIN – they have to make room in their bid, budget as part of a contract [$E_{28}$].

[309] **Dr S.** contends that MEDIN does not appear to be well known, unlike other portals such as the US Department of Commerce: National Oceanic and Atmospheric Administration (NOAA) [$S_{29}$]. NOAA is very well known, acknowledged by research users and a fruitful resource for oceanographic data – it is one of the first places many academics go for data [$S_{29}$]; *US Department of Commerce: National Oceanic and Atmospheric Administration (NOAA) Website* <http://www.noaa.gov/index.html> [accessed 9 August 2015].

what MEDIN is and where to go to it to find the data' [$W_{24}$].[310] However, it must be made clear that at the time of the interviews MEDIN was still in its early stages (for instance the bathymetry surveys were not yet released through the MEDIN portal) and this may have had an impact on the level of awareness.[311]

Although MEDIN provides open meetings and a newsletter [$N_{13}$], publicising MEDIN beyond these means is problematic.[312] As MEDIN is a UK government mechanism for marine environmental data re-usage, it has financial constraints and is unable to spend money on publicity events [$W_{24}$]. MEDIN is therefore reliant on gradually building a critical mass of research users over a longer period of time.

As is the case with many sectors and disciplines, the reliance of research users on search engines to locate data appears to contribute significantly to the underuse of the MEDIN portal [$W_{25}$]. There was a lack of consensus amongst the interview participants surrounding how the marine community were searching for data. It was suggested that: research users seeking public sector data would initially visit the data.gov.uk website, which would then direct them to MEDIN [$T_{24}$]; and, academic and other scientific researchers would go to MEDIN directly [$T_{24}$]. However, on the latter point **Dr S.** offered a dissenting opinion suggesting that those academics that are already familiar with marine environmental data providers would first visit a search engine to locate data, and those academics unacquainted with such data providers would first access a specialist portal such as MEDIN [$S_4$, $S_5$].

However, the interview participants agreed that using a discovery metadata portal such as MEDIN was more advantageous than searching for data via a search engine. The main reason appears to be that search engines only locate but do not discriminate between data sources as to quality and provenance, which may be variable.

---

[310] Although it is outside this thesis, **Mrs T.** further believes that there is a possible lack of awareness of MEDIN amongst private sector companies, which impedes the amount and range of marine environmental data deposited at the data archive centres [$T_{27}$]. **Mrs T.** thinks that the main contributors to MEDIN 'are mostly public sector […] there is a lot of private sector data […] that's not necessarily reaching *um* the MEDIN portal […] [such as] oil companies, *um* wind farm developers, utility companies, people laying pipelines and cables and so on which are relevant to the – the marine environment' [$T_{27}$]. This underuse and lack of awareness is partly because MEDIN is new, but MEDIN needs to directly target these individuals and organisations to secure their involvement [$T_{28}$].

[311] **Ms E.** highlights that at the time of the interview (2012) MEDIN is actively expanding by incorporating new thematic data archive centres, as shown within the primary source materials [$E_9$].

[312] A brief article about this PhD research was published in the MEDIN newsletter: Laura German, 'MEDIN Case Study: PhD Research Accepted by the 'ICT for Environmental Regulation Workshop'', *Marine Data News*, 24 June 2013 <http://medin.newsweaver.co.uk/marinedata/16had7wvxub> [accessed 9 August 2015].

The discovery portal by contrast not only facilitates a more advanced and tailored search but locates data held within accredited data archive centres.[313] To further signpost the MEDIN discovery portal to research users, the datasets held by the data archive centres need to individually appear on search engine results to increase awareness of the MEDIN portal [$W_{25}$].

From the primary source materials and the interviews, it is also clear that portal multiplicity is an emerging problem [$T_{23}$]. While MEDIN is actively addressing the past problems of marine environmental data duplication, it is now part of a new duplication issue: multiple discovery portals duplicating the signposting of marine environmental data. For instance, the bathymetric survey data held by the UKHO are (potentially) caught between five discovery portals [$N_{10}$, $T_{23}$]: MEDIN; data.gov.uk [$N_{10}$]; the European INSPIRE portal (which would harvest marine environmental datasets from data.gov.uk [$N_{10}$]); the European Marine Observation and Data Network (EModNet) pilot bathymetry portal (EModNet provides a number of other portal that potentially overlap with MEDIN) [$T_{23}$]; and another pan-European data portal that would overlap data.gov.uk [$T_{23}$].[314]

There is a potential contradiction of the principle: gather once and use many times. The existence of multiple portals has a (potentially) negative impact on data discovery by creating a portal proliferation, causing confusion over data versioning, and therefore diluting the awareness and eroding the confidence of research users [$N_{10}$]. Is it clear who holds the definitive dataset, and who is harvesting datasets? Are the portals harvesting the most up-to-date version of a dataset? As is the case with the bathymetric survey data, the UKHO is expected to hold these data and update them once, and portals will harvest the data from the UKHO.[315] Moreover, each portal has a slightly different

---

[313] **Mrs T.** clearly explains the benefits of using the MEDIN portal to discover data, as opposed to a generic search engine:

> […] The MEDIN *er* portal is managed, and therefore you do have some confidence that data that you're pointed to – it has some sort of authority associated with it. And hopefully, is reasonably up to date. *Er*, just using Google [or another search engine] would lead you to lots of – questionable data, *er* and certainly out-of-date data. [$T_4$]

[314] Refer to: 'Pilot portal for bathymetry', *European Marine Observation and Data Network (EMODnet) Website* <http://www.emodnet-hydrography.eu/> [accessed 9 August 2015].

Further to this, **Mrs T.** anticipates there will be 'closer integration' between MEDIN and data.gov.uk [$T_{23}$].

[315] **Mr N.** further explains this point:

> […] And it's more cost effective for a data provider or publisher, like ourselves [the UK Hydrographic Office], to make data available and updates to data once. Rather than having to make those updates available several times to different organisations. […] I think it's more likely

agenda for signposting data, for instance some international portals aim to provide further services and applications [$B_{23}$].

Portal multiplicity is problematic, as it contradicts the individual aims of each portal: to overcome duplicity and offer a single robust and authoritative platform for data discovery. The rationale behind MEDIN – gather once, and use many times – should be echoed and adopted by portal providers as: signpost once, and discover many times. Therefore, not only does data and provenance require management, but the portals used to signpost them.

## 4.4 MEDIN: interim conclusions

MEDIN has successfully built a sustainable and re-usable source of quality marine environmental data that can be shared across the marine sciences community (both within and outside the UK via INSPIRE). The key strengths of MEDIN are the employment of common standards, formats and practices that have been agreed and adopted by the marine sciences community.

Each marine environmental dataset is released under a data licence agreement, complete with a discovery metadata record through a user-friendly and authoritative metadata catalogue. The stakeholders are all clear about what management and re-usage purposes are authorised. The marine environmental data are quality checked and maintained by seven MEDIN accredited thematic data archive centres. This is an independent peer-review layer. As most of the data are openly released, these data are further open to public scrutiny. All data are released under open formats with no requirements for research users to purchase or learn new software.

In consequence, it is clear that MEDIN largely resolves the first major issue raised by Hwang and others, by effectively safeguarding diverse types of academic research data from multiple originators, contexts and sources for a wide set of users.

---

that we will aim for a – a more stream-lined situation where one update is managed – is – is provided, and then the various, different portals gather the data from – from what we've done. [$N_{10}$]

**Mr W.** adds to this by outlining the BODC's position:

> […] We [BODC] would generally prefer that people *erm* – come to us to actually access the data rather than the publisher. *Er*, just because *er* we've got the full history of the provenance of it and also *erm* any issues of versioning, and – and we know which is the definitive version as well. […] [$W_7$].

MEDIN's provenance metadata, legal, technological and socio-cultural frameworks are clearly proven as robust via evaluation of the primary source and interview materials. As a result, it is self-evident that through these frameworks MEDIN addresses many of the worst practice loopholes found in the case of Hwang and others. Moreover, there have been no instances of academic misconduct connected to MEDIN.

However, one weak point was revealed during the interviews. As with many models, MEDIN does not have the resources to ensure that research users are re-using data in accordance with the parameters of a data licence agreement. There is dependence on the user community to report any potential allegations of misconduct. As the case of Hwang and others showed, without the young South Korean academic community and use of online message boards to anonymously raise the alarm, the validity of the academic research data would not have been questioned. However, as MEDIN provides such a comprehensive provenance record of each dataset, this audit trail can be used to either support or oppose future allegations of scientific misconduct. Such robust provenance information was not available within the case of Hwang and others. Therefore, the reliance on good will and trust of the user community is a minor but potential limitation of this model.

In answer to the question: what makes for excellent quality academic research? – five key themes arise from this case study: (1) sustainability, (2) discoverability, (3) working towards a common understanding, (4) offering a good user experience, and (5) accreditation. These themes are now explained within this interim conclusion, and will be further evaluated as part of Chapters 7-8.

At the core of the MEDIN model is provision of a sustainable and re-usable resource for marine environmental data. This is aptly summed up by the crucial principle: gather data once and use many times. MEDIN was set up as a single authoritative mechanism not only to offer guidance on best practice, but to actively safeguard marine environmental data for re-use now and in the future.

Discoverability is also a principal component of MEDIN, as shown by the key role of the MEDIN discovery metadata. There is considerable political interest in this area as policy makers require the highest quality data for future decision making. As a result, many academics also write reports to be understood by a lay audience.

Moreover, academics are now re-using more existing data within their own data collection studies. Discovery metadata are also able to bring together associated marine environmental data that are dispersed over a number of thematic data archive centres.

Therefore, adding further value to the discovery of data by signposting the relationships between datasets either via theme, discipline or data gathered under a specific project or survey. For that reason, MEDIN is actively overcoming data silos through enhanced discovery.

MEDIN shows that working towards a common understanding of best practice across the marine environmental community has been paramount to its success. Changing organisational procedures can obviously be difficult, because people need to be shown the benefits of data re-usage and renewed means of best practice. MEDIN has managed to achieve this by building up a critical mass of active contributors, including: seven thematic data archive centres, fourteen sponsor organisations including NERC and around sixty bodies involved with MEDIN across governmental, commercial and academic sectors in the UK. Community involvement has been essential (in other words 'getting everyone in the same room') through a variety of working groups, encouraging contributors to influence and commit to the adoption of new practices. MEDIN's core team has a significant impact by co-ordinating these groups. Alongside the working groups, European standards for metadata such as INSPIRE have also helped the UK marine environmental community to work towards a common understanding through MEDIN. NERC data policy has also had an impact, as researchers are mandated to share their data where possible.

A degree of flexibility is an important component of MEDIN. For instance, while it would be beneficial from a licensing management perspective to reduce the overall types of licence in MEDIN, it is a concern that data suppliers would want unnecessarily to use the most restrictive licence. In consequence, although standardisation is important it is not an absolute requirement; in certain circumstances it can hinder re-usage.

The provision of a good user experience is vital for any research model to be effective. MEDIN's open community approach to creating its IT infrastructure and commitment to simplicity proves to be one of the key hallmarks of its usefulness. Research users do not have to purchase new software or learn new technical skills, and the discovery metadata catalogue is user-friendly. MEDIN also provide considerable support to the user community through its discovery metadata tool and helpline. The responsibility for providing crucial metadata lies with data originators. However, researchers do not have to be compliant with MEDIN discovery metadata standards on submission to a thematic data archive centre, the metadata can be in any format. To

what extent this level of service could be provided within other domains is questionable however.

Finally, accreditation is important, as each thematic data archive centre has to comply with twelve minimum standards of best practice, continue to meet those standards, and obtain approval from the executive team. All marine environmental data held by the thematic data archive centres undergo quality assurance and control procedures. However, while accreditation of data originators, suppliers and custodians is essential, research users are largely exempt from any verification procedures. For example, research user registration is not required and, as previously stated, re-usage operates on goodwill and trust.

In addition to the five key themes raised through the case study analysis, the interviews further highlighted three grey areas for further consideration within Chapters 7-8, namely: (1) the impact of funding on the overall effectiveness of MEDIN; (2) how portal multiplicity could potentially counteract MEDIN's aims to gather once and use many times; and, (3) the difficulties managing multiple licensing agreements.

The budget available to implement renewed best practice can be a critical issue for model sustainability. For instance, while MEDIN are aware of the benefits of linked data there was limited funding to implement these. Moreover, due to government funding constraints MEDIN was unable to advertise the model to attract a wider pool of users. As budget varies dramatically and is dependent on the specific type and circumstances of data acquisition, storage and dissemination – how can economic factors be better managed to reduce the risks to sustainability?

Secondly, model multiplicity not only jeopardises the sustainability of a model, it may also adversely impact on the discoverability of marine environmental data. Bathymetry surveys data are already experiencing this through five different portals attempting to do the same task. Multiplicity of depositories can seriously compromise the principal of gather once and use many times. The number of possible ways available to discover data further leads to the question: at what level do people discover data – do they begin with a search engine or go directly to a particular repository of choice?

Finally, MEDIN has to manage a large number and variety of data licence agreements, which can be problematic for rights management and clearance. The thematic data archive centres all operate different licensing frameworks. Moreover, as data custodians the thematic data archive centres store the majority of data in accordance with permissions granted by third parties, there is no one-size-fits-all

approach to data licensing. Furthermore, it appears that it would not only be impracticable for MEDIN to insist on the use of one standard licence, but unfeasible as MEDIN is not a legal organisation. How can data custodians, such as MEDIN, better manage this licensing complexity?

In summary, MEDIN demonstrates that it is possible to integrate a mixed previous data economy derived from multiple data originators, managers and users. Diversity is kept at the centre of MEDIN's interests, which allows it to manage data for many future uses not necessarily envisaged at the present time: gather once and use many times – in the future. The next chapter now turns to the issue of modelling best practice for academic research data generated via highly collaborative research environments.

# Chapter 5: eCrystals and LabTrove

The second case study is located in the highly collaborative discipline of crystallography. A lack of defining roles and contributions for the co-investigators in the case of Hwang and others (used as a motivating example of worst practice in Chapter 2) contributed to the erroneous data escaping detection through the peer-review process. For instance, one of the co-authors Gerald Schatten had not taken part in the practical experimentation and therefore was not in the position to be named as a senior co-author. Moreover, as a highly-regarded principal investigator Hwang should have overseen the research, but he later claimed that he was deceived by his junior researchers. This chapter therefore focuses on: how should academic research data generated as part of collaborative research be treated in order to balance a range of difficult communal and continuity problems?

The speed and volume in which crystallographic data are generated have increased significantly over the past number of decades. Around fifty years ago, a researcher would have spent the course of a PhD investigating around three crystal structures. Today, three crystal structures can be investigated in a single morning. Given this unprecedented pace of research, it is unsurprising that the majority of crystallographic data are unable to be published through peer-reviewed journals. This case study focuses therefore on two data management platforms, eCrystals and LabTrove, which have been established to increase the amount of chemistry data safeguarded for future re-usage.

eCrystals and LabTrove were developed by the Southampton Chemical Crystallography Group (SCCG) who are based within the department of chemistry at the University of Southampton. eCrystals is a forerunner to the digital open data movement, providing an openly accessible online data archive centre for crystallographic data outputs generated by the SCCG:

> eCrystals – Southampton is the archive for Crystal Structures generated by the Southampton Chemical Crystallography Group and the EPSRC UK National Crystallography Service. […][316]

Chemical crystallography offers accurate molecular measurements of crystal structures, and the SCCG utilise a scientific technique called single crystal X-ray diffraction,

[316] *eCrystals Website* <http://ecrystals.chem.soton.ac.uk/> [accessed 9 August 2015].

through the employment of four diffractometers to gather data about these measurements.[317] By obtaining these data within the laboratory, researchers are able to capture data about crystal structures of biological and chemical significance, such as for consequent utilisation within pharmaceuticals or as catalysts.[318]

As part of the National Crystallography Centre, the SCCG rely on data samples being sent to their laboratory from higher education institutions and other laboratories within the UK (and in some cases from other parts of the world). By comprising 1017 crystal structure data records produced from 1987 to 2011 by over 270 individuals (correct on 5 July 2015), eCrystals provides an extensive record of crystallographic data generation from a sizeable number of collaborators.[319] Therefore, this case study offers a more complex type of collaboration than experienced within Chapters 4 and 6. Not only are the SCCG a collaborative research team, but they are largely reliant on other collaborative research teams from across the UK to deliver crystallographic samples for analysis at the University of Southampton laboratory.

Any resulting data findings and research outputs must be jointly authored by the members of the SCCG involved and the external research team that initially discovered the crystallographic sample. All researchers must mutually agree which permissions and

---

[317] The Southampton Diffraction Centre provides a brief summary of single crystal x-ray diffraction:

> Single crystal X-ray diffraction is the ideal technique to elucidate the molecular structure of a crystalline material. Beyond the molecule, X-ray diffraction reveals interactions between the molecules as they assemble in the solid state.

'Small Molecule Diffraction', The Southampton Diffraction Centre, *University of Southampton Website* <http://www.southampton.ac.uk/sdc/small_mol_diffraction/index.page> [accessed 9 August 2015]. For further background information about crystallography refer to: 'Crystallography Collection', *Royal Institute Channel Website* <http://www.richannel.org/collections/2013/crystallography> [accessed 9 August 2015]; 'Crystallography timeline: Explore the history of one of the greatest innovations of the twentieth century'. *Royal Institute Channel Website* <http://www.rigb.org/our-history/history-of-research/crystallography-timeline> [accessed 9 August 2015]; *The British Crystallographic Association Website* <http://crystallography.org.uk/> [accessed 9 August 2015].

[318] The Chemical Crystallography Group offers a description of crystallography:

> Chemical Crystallography provides accurate and precise measurements of molecular dimensions in a way that no other science can begin to approach. […] Historically, chemical crystallographers have tended to concentrate on using single-crystal X-ray diffraction to determine the structure of what may be thought of as "small molecules"; the upper limit might typically be considered to be up to a few hundred non-hydrogen atoms. Chemical crystallographers study compounds which are both of chemical and biological interest, *e.g.* new synthetic chemicals, catalysts, pharmaceuticals, natural products ... the list is endless! […]

'About'. *The Chemical Crystallography Group Website* <http://ccg.crystallography.org.uk/about.shtml> [accessed 9 August 2015].

[319] 'Browse by person', *eCrystals Website* <http://ecrystals.chem.soton.ac.uk/view/people/> [accessed 9 August 2015].

restrictions must be placed on any resulting datasets. Collaboration does not always occur with known individuals and within the specific parameters and timings of a research project. Populating the eCrystals data archive is an ongoing process that involves many ad-hoc collaborations with laboratories across the UK and beyond. Therefore, these changeable research teams cause greater difficulties for balancing mutual permissions, communal quality control and joint authorship. However, it must be noted that eCrystals currently has five frequent users who (re-)use these data – a small quantity in comparison with the number of data originators that help to populate the archive.

LabTrove is electronic laboratory notebook software developed by the University of Southampton, which is openly available for download by anyone with access to the Web:

> […] The LabTrove application enables researchers to share their experimental plans, thoughts, observations and achievements with the wider online community in a secure, semantically rich and extensible manner. […] scientists will no longer have to print out data results to insert into conventional lab books; instead, results will be logically associated with the experiment and therefore accessible as required. Thus it is possible to pivot the material and view the data in a chronological diary form for example or data-centrically in terms of the scientific argument.[320]

eCrystals and LabTrove provide two different models of academic research data re-usage within the laboratory environment. However, this adds value to this case study as it offers direct access to various provenance metadata, legal, technological and socio-cultural frameworks all managing data generated via experiments.

By first examining the primary source materials, this case study investigates the provenance metadata, legal, technological and socio-cultural frameworks determined by eCrystals and LabTrove. This critical analysis highlights the key issues that require further clarification by the six interview participants. There follows an evaluation of the safeguards that facilitate the re-usage of data from multi-organisational, collaborative research teams. This chapter then considers to what extent these safeguards remedy the second major issue raised by the Hwang case, and how this model could resolve potential current and future cases of bad and worst practice.

---

[320] 'LabTrove, About us', *LabTrove Website* <http://www.labtrove.org/aboutus/> [accessed 9 August 2015].

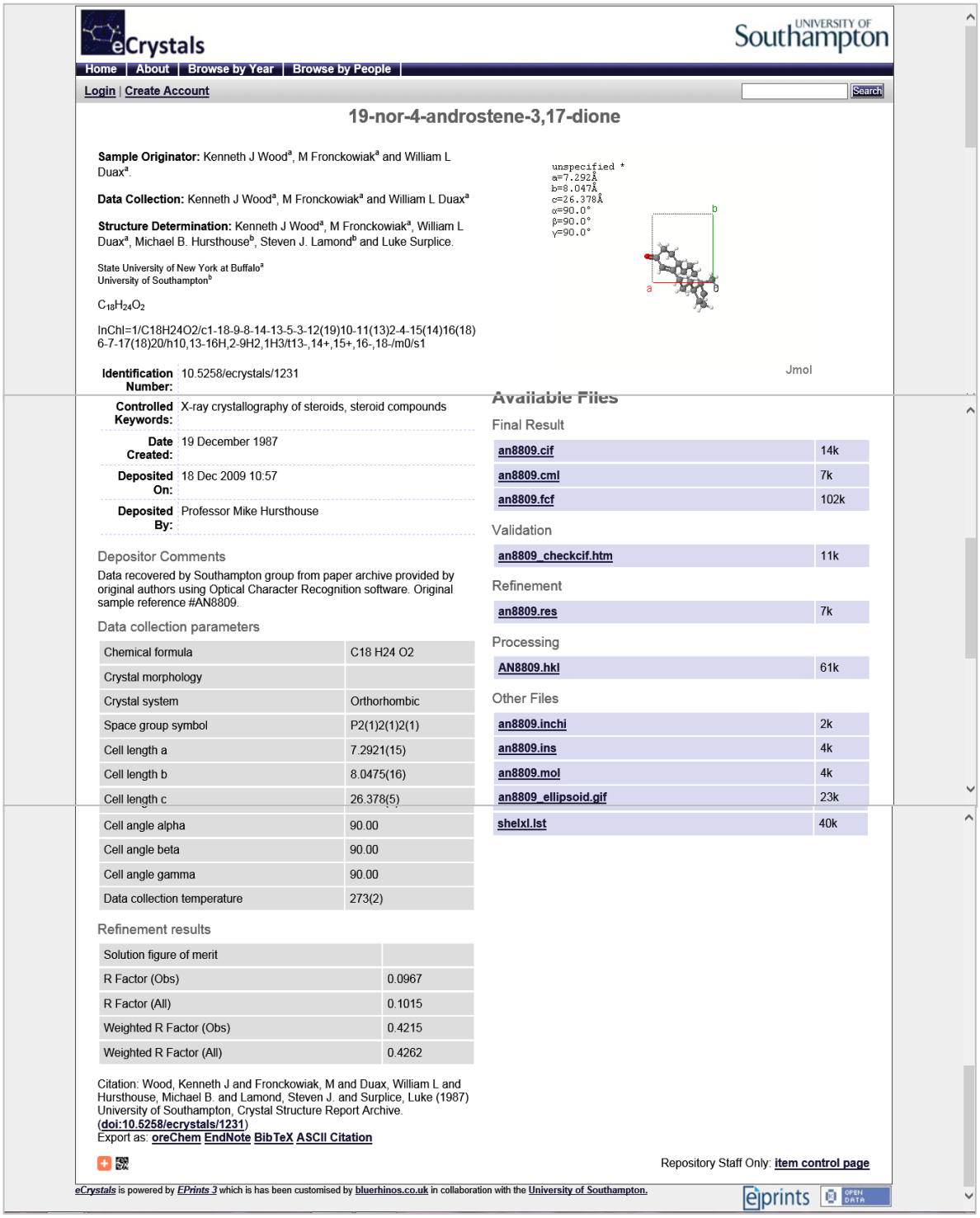



**Figure 3 Screen shot of '19-nor-4-androstene-3,17-dione', *eCrystals Website* <http://ecrystals.chem.soton.ac.uk/1231/> [accessed 29 October 2014].**

**Citation: University of Southampton, Crystal Structure Report Archive (2004). Image taken on 3 July 2013. Reproduced with permission from eCrystals, University of Southampton.**

## 5.1 eCrystals and LabTrove: primary source materials

### 5.1.1 eCrystals and LabTrove's rationale

eCrystals and LabTrove derive from a series of overlapping and associated research projects undertaken by the department of chemistry at the University of Southampton and other collaborators since 2001. Eight years after the worldwide release of the Web, the CombiChem project based at the University of Southampton was awarded an EPSRC grant (GR/R67729/01) to explore the re-usage of chemistry data in a digital age.[321] This grant ran from the 1 October 2001 to 31 March 2005 and focused on provenance:[322]

> The goal of the CombiChem project is to develop an e-science testbed that integrates existing structure and property data sources within a grid-based information-and knowledge-sharing environment. The service-based grid computing infrastructure extends to devices in the laboratory and involves enriched streams, (including multimedia and live metadata), full support for provenance and innovative techniques for automation throughout the environment.[323]

The principal investigator was Professor Frey from the department of chemistry who – with a number of co-investigators both from chemistry and the electronics and computer science departments – namely Professor M. Luck, Professor L. V. Moreau, Professor A. Welsh, Professor D. C. De Roure, Professor J. W. Essex, Professor S.M. Lewis, Professor M. Hursthouse and Professsor A. Orpen – developed CombeChem: a virtual research environment for combinational chemistry research.[324]

CombiChem was a pioneering open laboratory project. It appears to have been the initial spark for the eBank UK project that developed eCrystals and the Smart Research Framework (SRF) which produced LabTrove. This is now explained in the following sections.

---

[321] 'Structure-Property Mapping: Combination Chemistry & the grid (CombiChem)', Research grant, EPSRC Reference: GR/R67729/0, *Engineering and Physical Sciences Research Council (EPSRC) Website* <http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=GR/R67729/01> [accessed 9 August 2015].

[322] *Ibid.*

[323] *Ibid.*

[324] *CombeChem Website* <http://www.combechem.org/index.php> [accessed 9 August 2015].

#### 5.1.1.1 eCrystals and eBank UK

From 2003-7, eCrystals was developed during the interdisciplinary eBank UK project, which was 'led by [the United Kingdom Office for Library and Information Networking based at the University of Bath] UKOLN in partnership with the Intelligence, Agents & Multimedia Group, Department of Electronics & Computer Science and the Department of Chemistry, University of Southampton':[325]

> The eBank-UK Project (JISC-funded in three phases since September 2003), has investigated the feasibility of data repositories for the archiving and storage of crystal structure data, and the linking from primary data to other research outputs within the scholarly knowledge cycle. […] and constructed an institutional repository eCrystals that makes available the raw, derived and results data from a crystallographic experiment.[326]

eCrystals recognised the need for an interdisciplinary team to develop the data archive, a team which included: chemists, librarians, information scientists, and computer scientists.[327] eCrystals was established to address insufficient crystallographic data capture and re-usage within the crystallography community. Moreover this problem has enlarged over the past fifty years, as crystallographers experienced an increased rate and efficiency of crystallographic analysis leading to a growing amount of crystallographic data.

Crystallography was selected by eBank-UK as a test case for a digital chemistry archive, as it had been computational for many years – for example, the 'information associated with crystal data collection (raw data from the diffractometer, crystal coordinates and other structural parameters) was already available in digital format':[328]

---

[325] 'e-Bank UK: About the project', *The United Kingdom Office for Library and Information Networking (UKOLN) Website* <http://www.ukoln.ac.uk/projects/ebank-uk/>[accessed 9 August 2015]; 'About', *The United Kingdom Office for Library and Information Networking (UKOLN) Website* <http://www.ukoln.ac.uk/about/> [accessed 9 August 2015]; for further information about the e-Bank UK Project refer to: Liz Lyon, 'eBank UK: Building the Links Between Research Data, Scholarly Communication and Learning', *Ariadne,* 36 (2003) <http://www.ariadne.ac.uk/issue36/lyon/> [accessed 9 August 2015].

[326] Liz Lyon and others, 'Scaling Up: Towards a Federation of Crystallography Data Repositories', UK eBank Report, Version 1.0: Final (12 May 2008), p. 7 <http://www.ukoln.ac.uk/projects/ebank-uk/dissemination/Ebank3report/Ebank3report.pdf> [accessed 9 August 2015].

[327] Gráinne Conole, 'External evaluation of the eBank project', Final Independent Report (16 December 2006) p. 9 <http://www.ukoln.ac.uk/projects/ebank-uk/evaluation-report-dec-2006/evaluation-report-december-2006.pdf> [accessed 9 August 2015].

[328] Conole, p. 8.

> Focusing on x-ray crystallography as a subject area helped, as the workflow processes were already partially in electronic format and there was a degree of common agreement amongst crystallographers about file formats and work processes which meant that it was possible to provide a reasonable demonstrator of the potential across the work flow process from crystal data collection to publication.[329]

Therefore, eCrystals was able to utilise a pre-existing level of standardisation of best practice amongst the crystallography community.

#### 5.1.1.2 LabTrove and SRF

Initially developed around 2009, LabTrove aims to digitally record scientific experiments. However, it is not the first laboratory notebook developed by researchers within the department of chemistry at the University of Southampton. As part of the CombiChem project, a sub-project focused on creating an electronic laboratory notebook to be used by chemists at the University of Southampton – this was called the Smart Tea Project:[330]

> Smart Tea is about improving the information environment for chemists doing chemistry - within and beyond the lab. Smart Tea is about supporting chemists in the preparation, execution, analysis and dissemination of their experimental work. [/] When chemists run experiments, they create a great deal of information […] [/] For all its technical sophistication, the modern lab experiment is still recorded using the same tools as scientists have been using over the past 200 years: a bound paper lab book […] [/] is a poor mechanism for making the information stored in that book available to other scientists within the lab, or for that matter, to the same scientist after the experiment has been completed: if the scientist does not have the lab book to hand, the information is unavailable. [331]

Following on from the Smart Tea Project, LabTrove was developed to harness the benefit of sharing data from experiments on the Web. Like a blog, LabTrove users are able to record their experiments in their own personal style and choose the individuals with whom they share these data and information. LabTrove 'enables researchers to share their experimental plans, thoughts, observations and achievements with the wider online community in a secure, semantically rich and extensible manner'.[332]

---

[329] Conole, p. 9.

[330] *The Smart Tea Project Website* <http://www.smarttea.org/> [accessed 9 August 2015]; The Smart Tea Project Team consists: Jeremy Frey (Principal Director, Chemistry), David De Roure (Principal Investigator, Computer Science), Gareth Hughes, Hugo Mills, Terry Payne, m.c. schraefel (lower case deliberate) and Graham Smith.

[331] *The Smart Tea Project Website* <http://www.smarttea.org/> [accessed 9 August 2015].

[332] 'About us', *LabTrove Website* <http://www.labtrove.org/aboutus/> [accessed 9 August 2015].

LabTrove was developed as part of the SRF and involves a number of researchers directly involved with eCrystals:

> The Smart Research Framework (SRF) is a collection of three tools: LabTrove and Blog3, two alternative blogging platforms adapted for use as electronic laboratory notebooks (ELNs), and LabBroker, a tool for piping data from instruments into ELNs. This project developed these tools into cloud-based services for the UK higher education sector.[333]

LabTrove is available for open download, and although still in its infancy there are international examples of its utilisation. For instance, LabTrove is used by the University of Sydney as part of its Open Malaria Project.[334]

### 5.1.2 eCrystals: overview

The department of chemistry at the University of Southampton is a world-leading facility for single crystal X-ray diffraction.[335] The Southampton Chemical Crystallography Group (SCCG) based at the University receive crystal structures from laboratories across the UK to analyse by using specialist equipment – diffractometers.[336] eCrystals is a data archive for crystallographic data generated via this process of analysis within the laboratory:

> […] The information contained within each entry of this archive is all the fundamental and derived data resulting from a single crystal X-ray structure determination, but excluding the raw images. […] Any reader wishing to have access to the raw images is welcome to contact us, and we will make arrangements for these to be made available.[337]

Through the eCrystals website, crystallographic data are made freely available to anyone with Web access, which enables research users to 'a) assess the validity of the

---

[333] Alex Ball, 'Smart Research Framework', 1 April 2011, *UKOLN Informatic Group, University of Bath Website* <http://irg.ukoln.ac.uk/2011/04/01/smart-research-framework/> [accessed 9 August 2015].

[334] 'Open Malaria Project', Open Source Malaria Website <http://opensourcemalaria.org/> [accessed 9 August 2015]; *Open Malaria Project Website* <http://malaria.ourexperiment.org/> [accessed 9 August 2015]; Jeremy Burrows, 'Advancing antimalarial drug research through open source initiatives', *Guardian*, 24 July 2013 <http://www.theguardian.com/global-development-professionals-network/2013/jul/24/open-source-drug-discovery-research> [accessed 9 August 2015].

[335] 'Small Molecule Diffraction', The Southampton Diffraction Centre, *University of Southampton Website* <http://www.southampton.ac.uk/sdc/small_mol_diffraction/index.page?> [accessed 9 August 2015].

[336] 'Equipment', The Southampton Diffraction Centre, *University of Southampton Website* <http://www.southampton.ac.uk/sdc/small_mol_diffraction/equipment.page> [accessed 9 August 2015].

[337] *eCrystals Website* <http://ecrystals.chem.soton.ac.uk/> [accessed 9 August 2015].

dataset or b) repeat the experiment or c) use the data for further studies'.[338] The process of depositing crystallographic data into eCrystals is explained by Lyons and others:

> Following the creation of a completed crystal structure, data is uploaded into a data repository and additional metadata (chemical & bibliographic), to Dublin Core standards, is associated with the dataset. This approach allows rapid release of crystal structure data into the public domain, but can also provide mechanisms for value added services that allow discovery of the data for further studies and reuse, whilst ownership of the data is retained by the creator. [339]

Similar to MEDIN, eCrystals aims to make these data as accessible to as many research users as possible.

As with MEDIN, provenance metadata has a pivotal role within eCrystals. Crystallographic data and metadata are presented in a standardised layout, as the eCrystals website describes:

> […] An individual entry consists of three parts:
> 1. Core bibliographic data, such as authors, affiliation and a number of chemical identifiers,
> 2. Data collection parameters that allow the reader to assess at a glance certain aspects of the crystallographic dataset,
> 3. Files available for download. These files are: visualisations of the raw data (.jpg), the raw data itself (.hkl), experimental conditions(.htm), outputs from stages of the structure determination (_xs.lst, _xl.lst & .res), the final structural result (.cif & .cml) and the validation report of the derived structure (_checkcif.htm). […][340]

eCrystals has a robust provenance metadata framework, which offers a rich and sizeable number of metadata fields, including: sample originators; data collection; structure determination; depositor; and comments of the depositor that capture the individuals' names, roles and respective institutions. Alongside the name of the crystal structure and its chemical formula, eCrystals also has metadata fields to record the identification number, controlled key words, date of creation, date of deposit and data collection parameters (such as the data collection temperature). Moreover, there are metadata fields to capture the refinement of the result, and various files containing data across the lifespan of the experiment including: final result, validation, refinement, processing and other.

---

[338] 'About the Repository', *eCrystals Website* <http://ecrystals.chem.soton.ac.uk/information.html> [accessed 9 August 2015].

[339] Liz Lyon and others, 'Scaling Up: Towards a Federation of Crystallography Data Repositories', p. 7.

[340] 'About the Repository', *eCrystals Website* <http://ecrystals.chem.soton.ac.uk/information.html> [accessed 9 August 2015].

It is evident from the eCrystals website that its provenance metadata appears to be very structured and is largely repeated for each crystal structure archived. This detailed array of provenance metadata fields (that are available on a single webpage) provides a transparent record of research – and showcases a working open laboratory. How this metadata framework works in practice will be raised with the interview participants. As LabTrove is a personal account of research, it will further be important to not only find out more information about its provenance metadata framework, but compare it to the very rigid framework employed by eCrystals.

### 5.1.3 LabTrove: overview

LabTrove enriches this case study by providing an alternative model to eCrystals by providing further provenance metadata, legal, technological and socio-cultural frameworks and issues for consideration. It adds value as although it developed from the same series of projects as eCrystals, it offers another method to confront some of the data sharing issues faced by many laboratory-based researchers. In contrast to eCrystals where each dataset stored is subject to fixed constraints, LabTrove enables the researcher to specify their own parameters. Therefore, as an electronic laboratory notebook, it is obvious that it is difficult to precisely evaluate its provenance metadata, legal, technological and socio-cultural frameworks. This is because, the individual researcher controls what data are posted, which descriptive tags and formats are used, whether to openly release or embargo data, and what licence to apply to their data. However, LabTrove does openly release documentation to inform research users about its capabilities.[341]

### 5.1.4 eCrystals' legal framework

eCrystals utilises two open licences provided by Open Data Commons (launched in 2008) which are commonly known amongst the scientific community: (1) Open Data Commons Attribution License, and (2) the Database Contents License.[342] A brief overview of Open Data Commons is provided in Chapter 2, section 2.3.3. The following statement is given on the homepage of the eCrystals website:

---

[341] *LabTrove Documentation Wiki* <http://docs.labtrove.org/dev/lt/Main_Page> [accessed 9 August 2015].
[342] *eCrystals Website* <http://ecrystals.chem.soton.ac.uk/> [accessed 9 August 2015].

> This eCrystals repository is made available under the Open Data Commons Attribution License: http://opendatacommons.org/licenses/by/draft. Any rights in individual contents of the database are licensed under the Database Contents License: http://opendatacommons.org/licenses/dbcl/1.0/[343]

This homepage notice is useful, because it immediately signposts the conditions of re-usage to research users before they begin to search for crystallographic data. The Open Data Commons Attribution License: Summary Document permits research users:

- To Share: To copy, distribute and use the database.
- To Create: To produce works from the database.
- To Adapt: To modify, transform and build upon the database. (Bullet points from original webpage).[344]

However, these acts are only authorised where the research user attributes the database and flags the original data licence agreement to subsequent research users – otherwise known as an Attribute Share-a-like License.

> Licensor grants to You a worldwide, royalty-free, non-exclusive, perpetual, irrevocable copyright license to do any act that is restricted by copyright over anything within the Contents, whether in the original medium or any other. These rights explicitly include commercial use, and do not exclude any field of endeavour. These rights include, without limitation, the right to sublicense the work.[345]

The Database Contents License (DbCL) v1.0 is '[2.1] a worldwide, royalty-free, non-exclusive, perpetual, irrevocable copyright license to do any act that is restricted by copyright over anything within the Contents, whether in the original medium or any other […] [2.3] […] The Licensor takes the position that factual information is not covered by copyright'.[346] In consequence, the crystallographic data openly released through eCrystals permits a wide range of re-uses subject to attribution.[347] While eCrystals aims to make its data openly accessible, eCrystals actively negotiates with its joint data owners.

---

[343] *eCrystals Website* <http://ecrystals.chem.soton.ac.uk/> [accessed 9 August 2015].

[344] 'ODC Attribution Summary', *Open Data Commons Website* <http://opendatacommons.org/licenses/by/summary/> [accessed 9 August 2015].

[345] 'Database Contents License (DbCL) v1.0', *Open Data Commons Website* <http://opendatacommons.org/licenses/dbcl/1.0/> [accessed 9 August 2015].

[346] *Ibid.*

[347] 'Rights For eCrystals – University of Southampton', *eCrystals Website* <http://ecrystals.chem.soton.ac.uk/rights.html> [accessed 9 August 2015].

### 5.1.5 eCrystals' technological framework

eCrystals utilises existing open standards, including Jmol which is an open-source Java viewer found on each record to display the crystal structures as 3D images.[348] In contrast to MEDIN where a past lack of standard formats caused considerable difficulties for data sharing, eCrystals was able to use two established and community adopted standards: Crystallographic Information File (CIF) and Chemical Mark-up Language (CML). In 1991, CIF was developed as a universal exchange file, which enabled crystallographers to share data in a standardised format.[349] Moreover, in 1995, CML was created as the XML standard for chemistry by Peter Murray-Rust and Henry Rzepa (a point also highlighted in Chapter 2, section 2.3.1).[350] As a result, it appears that chemistry has an advantage over other disciplines with its established and common formats.

The eCrystals data archive is based on an adaptation of EPrints 3 software, which was developed for building publication repositories compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI).[351] However, to what extent this software extends to academic research data rather than publications is an issue requiring further exploration during the semi-structured interviews. There is limited

---

[348] 'Jmol: an open-source Java viewer for chemical structures in 3D', *Jmol Website* <http://jmol.sourceforge.net/> [accessed 9 August 2015].

[349] Sydney R. Hall, Frank H. Allen, and I. David Brown, 'The Crystallographic Information File (CIF): A New Standard Archive File for Crystallography', *Acta Crystallographica,* A47, 655-685 (1991) <http://www.iucr.org/__data/iucr/cif/standard/cifstd1.html> [accessed 9 August 2015]; 'Information about CIF Format Required', *The Royal Society of Chemistry Website* <http://www.rsc.org/Publishing/Journals/guidelines/AuthorGuidelines/AuthoringTools/CIFDataImporter/CIFFormatForCIfDataImporter.asp> [accessed 9 August 2015]; 'CIF', *International Union of Crystallography Website* <http://www.iucr.org/resources/cif> [accessed 9 August 2015]; 'CIF – Crystallographic Information Framework', *Digital Curation Centre (DCC) Website* <http://www.dcc.ac.uk/resources/metadata-standards/cif-crystallographic-information-framework> [accessed 9 August 2015].

[350] *Chemical Mark-up Language (CML) Website* <http://www.xml-cml.org/> [accessed 9 August 2015]. Refer to the following webpage for a number of useful articles by Peter Murray-Rust, Henry Rzepa and others: 'Chemical Markup Language – publications', *Chemical Mark-up Language – CML Website* <http://www.xml-cml.org/documentation/biblio.html> [accessed 9 August 2015].

[351] *EPrints Website* <http://www.eprints.org/software/> [accessed 9 August 2015]; The Open Archives Initiative Protocol for Metadata Harvesting', Protocol Version 2.0 of 2002-06-14. Document Version 2008-12-07T20:42:00Z, *The Open Archives Initiative (OAI) Website* <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm> [accessed 9 August 2015].

information on LabTrove's technological infrastructure; however, unlike eCrystals, the LabTrove system is data centric.[352]

eCrystals and LabTrove both utilise Digital Object Identifiers (DOIs) due to their involvement with the JISC funded DataPool Project at the University of Southampton (2011-2013). [353] eCrystals and LabTrove were used to trial DataCite – a service for minting DOIs – which is employed by the British Library.[354] Their participation in the project is an example of how these models are striving to continually improve the data sharing experience, and are a recognised and important feature of the University of Southampton's archival landscape.

### 5.1.6 eCrystals' socio-cultural framework

In contrast to MEDIN, data verification is largely at the user level, as stated on the eCrystals website:

> […] The results have not been externally refereed, but the information supplied should enable any reader to check the reliability and validity directly, since all the files provided are freely available for download. Should any error be detected, we would appreciate receiving suitable comments, and we will make any necessary amendments, and include a note to that effect. […][355]

However, each dataset has a data quality file attached complete with traffic light warning system, as Wendy A. Warr states:

> Quality control is applied to the archive to ensure that, as far as possible, the user gets a full and correct record. The Crystallographic Information File (CIF) is checked (a checkCIF feature is incorporated in a data manipulation toolbox), value-added data are handled, formats can be converted, and all associated metadata are stored.[356]

---

[352] 'LabTrove, About us', *LabTrove Website* <http://www.labtrove.org/aboutus/> [accessed 9 August 2015].

[353] *Jisc DataPool Project Website* <http://datapool.soton.ac.uk/about/> [accessed 9 August 2015].

[354] Steve Hitchcock, 'Trialling DataCite for chemistry lab notebooks and repository data services', 18 December 2012, *Jisc DataPool Project Website* <http://datapool.soton.ac.uk/2012/12/18/trialling-datacite-for-chemistry-lab-notebooks-and-repository-data-services/> [accessed 9 August 2015].

[355] *eCrystals Website* <http://ecrystals.chem.soton.ac.uk/> [accessed 9 August 2015].

[356] Wendy A. Warr, 'Digital Repositories Supporting eResearch: Exploring the eCrystals Federation Model', *Ebank/R4l/Spectra Joint Consultation Workshop*, Location: London, Date: 20 October 2006 (December 2006) (p. 4) <http://www.ukoln.ac.uk/projects/ebank-uk/workshops/eBank-SPECTRa-R4L-workshop/eBank-SPECTRa-R4L-workshop.pdf> [accessed 9 August 2015].

It is important that the extent to which eCrystals' socio-cultural frameworks operates and which issues arise are further explored during the semi-structured interviews.

## 5.2 eCrystals and LabTrove: people selected for interview

**Dr A.** is a researcher within the SCCG at the University of Southampton, and the UK's National Crystallography Service (NCS). She was also part of project that tested the functionality of LabTrove as a tool for use within scientific education. [$A_1$] **Dr A.** was selected principally for her data management expertise.

**Mr C.** is a research data manager within Library Services who is involved with research data management policy and strategy at the University of Southampton. Alongside academic researchers from chemistry, he is an investigator on the DataPool Project, which involves eCrystals as a test case for DOI minting. [$C_1$] **Mr C.** was selected principally for his data policy expertise.

**Dr G.** is an academic within the SCCG at the University of Southampton and the NCS. He is directly involved with the development of eCrystals and LabTrove from their creation to present, and therefore has an overview of both projects. [$G_1$] **Dr G.** was selected principally for his data management expertise.

**Mr H.** is a legal advisor within Legal Services at the University of Southampton who deals with intellectual property law and general contract law issues, amongst other legal problems at the University. [$H_1$] **Mr H.** was selected principally for his legal expertise.

**Miss J.** is a crystallography PhD candidate within SCCG at the University of Southampton. She was also part of project that tested the functionality of LabTrove as a tool for use within scientific education. She utilises LabTrove to keep records of her research projects. [$J_1$] **Miss J.** was selected principally for her academic research expertise.

**Ms R.** is a software engineer at the University of Southampton. She was selected due to her technological expertise and direct involvement with eCrystals. She worked with the chemistry department at the University of Southampton to develop the first implementation of eCrystals. Also in the past, she was directly involved with the development of ePrints Soton. [$R_1$] **Ms R.** was selected principally for her technological expertise.

**Dr A., Mr C., Dr G., Mr H., Miss J.** and **Ms R.** were interviewed during 2012-3. The interviews were approved by the University of Southampton's Management Ethics Committee and adhered to the planned methodology (see Chapter 3 for ethics notice and other information).

## 5.3 eCrystals and LabTrove: interview materials

### 5.3.1 eCrystals and LabTrove's provenance metadata issues

#### 5.3.1.1 Structured and unstructured

As with MEDIN, provenance metadata forms an integral part of both eCrystals and LabTrove. The interview participants consider provenance as important [$C_{12}$, $G_{11}$, $H_{13}$, $R_8$].[357] The provenance metadata framework used by eCrystals is formulaic, and all data records and metadata are structured in the same way on a single web page [$G_4$, $J_{11}$, $R_1$]. **Ms R.** elaborates on the standardised context which facilitates the creation of such systematic provenance metadata:

> […] One of the things that's interesting about the eCrystals is that it's *um* very, very standardised, shaped data. Unlike a lot of fields where the data's *um* very much dependent on the researcher – *um* the actual analysis […] from that lab followed – almost everything followed exactly the same shape of the initial […] information gathered from the machine, the processing steps, the final outputs. *Um* so it's very normalised compared with what you'd expect in some fields. [$R_1$]

The metadata framework [$J_{11}$] employed by eCrystals is also retrospective where an ontology is applied to all the data files on the system to generate a provenance trail [$G_{11}$].[358]

While the type of provenance metadata framework employed by eCrystals was evident from the primary source materials, there was very limited information available with regard to LabTrove's frameworks. This is because LabTrove users can decide on

---

[357] **Mr C.** emphasises eCrystals's strong provenance metadata framework:

> […] You don't really promote re-use unless people can understand and re-use the data effectively. So what additional documentation do you need to provide, what metadata do you need to provide to describe your data? These are all things where our colleagues in eCrystals have a good background in providing that information […] [$C_2$]

[358] There are plans to semantically encode the metadata in semantic XML format or RDF [$J_{11}$]. **Miss J.** is aware that there is a DOI schema available for semantically encoded data that will be potentially utilised [$J_{11}$]. **Mr C.** anticipates that the future Web will become a moving space where linked data will be important to describe relationships that are continually moving [$C_{18}$, $C_{19}$].

the scope of the provenance metadata that suits their personal preferences. Therefore in in comparison to eCrystals, LabTrove's provenance framework is described as nebulous and unstructured [$G_4$, $J_{11}$]. LabTrove allows researchers to tag their electronic laboratory notebook blog posts by selecting the most appropriate key terms, and adding new ones where required [$G_{11}$, $J_{11}$]. Blog posts are joined together organically in a many-to-many relationship to other posts tagged with the same key terms, therefore (unlike eCrystals and MEDIN) the data and provenance metadata records are dissimilar [$G_4$, $J_{11}$]. This metadata diversity causes considerable data discoverability and re-usage issues as **Dr G.** explains:

> […] How the links are all set up, depends on someone's personal preference. And, being as this is a personal diary, *um* you know, it has to be something that *er* the individual feels happy with […] and it gives maximum benefit for them. So when it comes around to others re-using it, you kind of have to sit there and work out: 'okay, this is how they've linked this, and this is why they've done in this way'. *Um* – so a human can come along *er* and work that out, *er* it's slightly harder for a machine […] [$G_4$]

As a result, in contrast to eCrystals which utilises a number of structured metadata fields, LabTrove has no such requirements, which can make it difficult for researchers not only to locate all the provenance information but follow the scientific story [$G_4$]. The LabTrove developers are currently investigating ways in which these unstructured metadata can be made machine-understandable, and therefore facilitate automated monitoring and discovery for research users [$G_4$]. For instance, alongside LabTrove the developers have produced an electronic laboratory notebook where each keystroke is recorded and expressed in machine-understandable forms [$G_{11}$]. Furthermore, the developers are also investigating which metadata fields need to be included to make data more discoverable through search engines [$G_4$].

Following on from this point, **Ms R.** provides a broader perspective as she foresees that the next generation of laboratory equipment will increase the production of provenance metadata, and further standards would need to be built around capturing metadata through this equipment [$R_{12}$].[359] In the event that more automated provenance metadata capture systems are used and provenance metadata becomes increasingly machine-understandable, meta-metadata (a record of how the provenance metadata was

[359] A large number of people with digital cameras will already be familiar with automatically generated provenance metadata (for instance the GPS location and shutter speed) [$R_{12}$]. However, depending on whether the digital camera has been set up properly, the wrong date and time can be automatically attached to all photographs uploaded onto a computer [$R_{12}$].

captured) may be required to ensure that the machine is functioning properly and has attached the correct provenance record to the right dataset [$R_{12}$]. Moreover, **Ms R.** maintains that this could lead to a multitude of repositories holding a number of different types of metadata [$R_{12}$].

There is a need to accommodate both structure and flexibility within academic research data management (as also highlighted by MEDIN) [$G_{11}$, $G_{12}$, $R_{12}$]. This is shown by the two very different provenance metadata frameworks employed by eCrystals and LabTrove. The level of standardisation varies within chemistry [$G_{12}$]. **Ms R.** contends that researchers require space for innovation and therefore one universal provenance metadata system that applies to all data types is not an option – there could be 'thousands of different formats' [$R_{12}$] for provenance metadata. Ideally, the data originators should be in control of choosing the provenance system, type and amount of metadata generated [$R_{12}$]. Whereas, data managers would be responsible – in a role as gatekeepers – for the institutional/bibliographic metadata fields [$R_{12}$]. This leads to the following questions: to what extent should provenance metadata fields be pre-determined and standardised? Does the level of prescriptivism depend on the main purpose of the academic model and type of data? For instance, structured metadata conform to the formulaic nature of crystallographic experiments, whereas unstructured metadata suit the personal nature of an electronic laboratory notebook such as LabTrove.

Provenance metadata are not only crucial for recording legal information such as the copyright owner [$H_{13}$], but for scrutinising the quality of a dataset [$R_{13}$].[360] However, provenance metadata cannot actively prevent or detect academic misconduct, as **Ms R.** states:

> [… ] I think quality assurance […] it's more about tracing back what happened. So knowing […] if you discover that *um* a lot of information that we've got errors in a whole bunch of our data and it turns out they all came from *urm* Microscope B, and all the ones from Microscope A are fine – we've learned an important thing about Microscope B; or, possibly Researcher B. […] but I don't think you're going to catch quality errors in […] the process. […] [$R_{13}$]

Therefore while provenance metadata are unlikely to proactively prevent genuine errors and/or academic misconduct, it can be used retrospectively to locate the source of such issues.

[360] **Mr H.** contends that it is important that the University is aware of the following points: where and how the data originated; who created the dataset; and, who employs the data originators [$H_7$].

For some disciplines, such as crystallography, there is a need to distinguish between multiple copies and versions of a dataset, as **Miss J.** states:

> […] publishing several different versions of the same dataset – *um* that actually to my mind that is a benefit not a problem […] there are circumstances under which sometimes you will process a dataset to the best of your ability *um* and […] you might publish it, as I say, an incomplete dataset. […] and then you process it again later using a different piece of software, or with more experience behind you, or any number of circumstances. And when you do that, you produce a better or a different result. Now actually you might want to publish both results as an illustration of the fact that your new software is more advanced in its ability to refine the data structure. […] I don't see that kind of data duplication as a problem particularly if it's stored with metadata that allows you to disambiguate the two entries. […] [$J_{19}$]

**Miss J.** further explains how provenance metadata is able to prevent confusion where datasets are intentionally duplicated:

> […] publishing the same crystal structure gathered in two different locations twice, *um* that is a problem, but […] if the proper automated systems were put in place to check for data duplication, then I foresee that will not necessarily be a problem. And again it could be a benefit in that someone using a – a different diffractometer they're on a different side of the world – there's always that idea of scientific validation by repetition from a different, disassociated research group. […] [$J_{19}$]

Provenance metadata needs to be able to keep pace with the continually changing definitive version of a dataset as, for example, data analysis techniques develop and/or a distinction needs to be drawn between different researchers (re-)using the same dataset, either for scientific repetition or to apply different research methodologies.

**Dr G.** raises an important question: is provenance metadata more than just tracking what researchers have done in the course of their research [$G_{11}$]? There is a difference between institutional/bibliographic metadata, and scientific (/academic) metadata [$G_{11}$, $R_{12}$] that captures the conditions of an experiment or semi-structured interview for instance. Provenance metadata have a number of layers which may require different sets of expertise to be completed accurately [$R_{12}$]. It is the responsibility of the researcher to complete the necessary academic level metadata – the data manager/archivist is unlikely to have the specific disciplinary knowledge about methodologies, experiments, theories, applications, processes and/or systems and will mainly be concerned with the institutional/bibliographic metadata [$R_{12}$].[361] For this

[361] **Ms R.** states:

> […] Ensuring the provenance of data is important. To say *um*, you know, we correctly record who submitted this information and […] maybe verifying that any files of a specific format are

reason, **Mr H.** is concerned about a reliance on the depositor who may have limited knowledge of the legal rights completing the legal metadata fields [$H_{13}$]. eCrystals does not appear to have a legal provenance metadata field for the reason that it only utilises two data licences; however this is an area that could be strengthened. It is not only data and provenance management that is vital, but the effective management of intellectual property rights [$H_7$].

Complex metadata with a number of layers may require more time and effort to maintain. **Mr C.** explains how the profile of provenance metadata must be raised as an integral aspect of research:

> […] the more data [that] becomes complex, I think […] the more you want to be able to really be absolutely nailed on about provenance of all the – the elements of your dataset. *Um* but equally, I think the more metadata [that] gets complex, the harder it is to achieve, because ultimately *er* we work in a very busy environment, […] complex metadata becomes almost unaffordable unless the culture changes to really be positive about this being part of the time that is taken up in the research process […] [$C_{12}$]

A final provenance metadata issue was raised by **Mr H.** who asked: how easy is it to de-link provenance metadata (and a licensing agreement) from a dataset [$H_{11}$, $H_{13}$]? This is an important question, as it further demonstrates that detailed and accurate provenance metadata are only useful in context. In the event that provenance metadata are de-linked and therefore creates an orphan dataset, that dataset will not be as re-usable to subsequent research users. Therefore, how to securely embed provenance metadata within a dataset (for instance water-marking) is an area for further investigation.

### 5.3.2 eCrystals and LabTrove's legal issues

As with MEDIN, eCrystals and LabTrove operate within the existing legal framework without difficulty on a daily basis.[362] For instance, **Dr A.** knows of no licensing or

---

> correctly formatted. […] [The University] enforce ethics […] procedures and so forth, but I don't think the University should be *um* reviewing each file individually […] that's peer review that's not *um* something the University itself could be doing. But I do think it's important to keep records so [...] if there are problems, you can then go and review. […] [$R_8$]

[362] Due to its historic development and subsequent extensions to different type of works, the multiple layers of different rights, database rights and contract rights makes copyright law fragmented and confusing [$H_{17}$]. **Mr H.** does not think that copyright law 'facilitates data reuse at all' [$H_{17}$], and further contends: 'there's always the argument that the law takes too long to catch up to what's happening in practice' [$H_{17}$]. **Mr H.** also states: 'case law helps us to clarify […] the meaning of adverse provisions in the [Copyright, Designs and Patents] Act' [$H_{12}$]. Without this case law, legal advisors offer advice based on their interpretation of the law [$H_{12}$]. As academic research data re-usage increases to fulfil funding bodies' obligations, **Mr H.** suspects case law may grow [$H_{12}$].

other legal issues associated with eCrystals or LabTrove [$A_9$, $A_{10}$], and **Mr H.** questions whether copyright law is truly a hindrance in practice [$H_{27}$]. However, it is uncertain whether this lack of legal difficulty is due to the limited use being made of both models. eCrystals has a small core user-base of around five researchers, most of whom are known to each other [$A_8$, $A_{16}$, $A_{34}$, $J_5$]. This limited use makes it more difficult to judge the effectiveness of its legal framework, although it appears that (in theory) eCrystals has a robust legal framework utilising two unrestrictive licences – the Open Data Commons Attribution License and the Database Contents License – that are familiar within the scientific community.[363]

#### 5.3.2.1 Open licensing

As it appears to be the duty of the individual to appropriately license their data through LabTrove, there was no cohesive legal framework to assess. However, it is likely that research users would opt to utilise open licences, such as Creative Commons and the Open Knowledge Foundation licences utilised by eCrystals; there is an increased use of these licences across higher education institutions [$H_{11}$]. For instance, **Miss J.** personally favours open licensing – in particular the Creative Commons BYSA licence, because a dataset has to be shared under the same original terms with an attribution statement [$J_9$]. This prevents individuals re-using a dataset and publishing it within a proprietary journal where subsequent research users may be unable to get access to it [$J_9$]. However, such share-a-like licences are not a perfect solution, as **Ms R.** explains:

> […] Share-a-like again is a tricky one, because […] you can't combine share-a-like work with something with another restriction on it. So *um* … I think … what would be useful is if we had some *er* legal guidance on it really. As in *er* someone said: 'well this case this happened'. But […] everyone's talking theoreticals [sic], because […] no-one's ever, you know, gone to court over it […] [$R_{15}$]

As with the MEDIN case study where it was stated that it was for the government to decide the appropriate licences for its organisations, this sentiment is reflected by **Ms R.** who argues that it is for research councils to determine the types of licences that should be utilised by researchers [$R_{15}$]. Moreover, the University's in-house legal services department is only a small team of individuals providing advice to a large organisation on consultation [$H_2$] without the resources or time to check the legal issues pertaining to

[363] **Ms R.** has faced problems where some extremely risk-adverse organisations are reluctant to license data, as they do not want to commit to mistaken terms and conditions [$R_{15}$].

each dataset [$H_2$, $H_7$], and therefore cannot undertake this responsibility in their present situation.[364] However, there are plans within Research and Innovation Services at the University to create contracts for data sharing [$H_8$].[365] There are also more general concerns that researchers have limited legal awareness and are often averse to asking for advice from legal professionals [$H_7$, $J_{10}$]; therefore perhaps a list of recommended licences could be beneficial as a starting point.[366]

During the interviews, **Mr H.**, **Miss J.** and **Ms. R.** debated the advantages and disadvantages of employing open licensing systems, such as Creative Commons [$H_{11}$, $J_9$, $R_{15}$]. While the choice and simplified style (such as the icons, plain language, and licence summaries) help highlight legal rights and obligations to researchers, there were concerns over whether these licences give researchers a rich enough legal overview as such licences only offer a basic summary of the law [$H_{11}$]. Moreover, open licences that restrict commercial re-usage can cause confusion for some researchers working within higher education [$R_{15}$]: does re-using a dataset to help a funding bid application constitute a commercial activity, or working with an industrial partner? There also appears to be a school of thought that researchers should favour the Creative Commons Zero licence – where the copyright holder waives all rights including attribution – as

---

[364] The general role of Legal Services is discussed at length by **Mr H.**, a key concern is that Legal Services are not consulted very often – this dialogue needs to be improved [$H_3$].

[365] **Mr H.** states:

> […] The contracts can be drafted better and can actually be used to try and – fill a position which is otherwise unclear by the various copyright laws, and so on, of the various jurisdictions who may be involved in that collaboration. [$H_9$]

[366] While eCrystals and the LabTrove research development team do not have a specific legal member, they have direct access to legal advice, if and when required, from Research and Innovation Services, Library Services and/or the Legal Services at the University of Southampton [$H_3$]. (This Chapter includes specialist insights from Library Services (**Mr C.**) and Legal Services (**Mr H.**); refer to Chapter 6 for data gathered from **Mr M.** – who is a legal advisor with Research and Innovation Services at the University of Southampton.) Although Legal Services aim to outline the legal issues and 'hopefully come up with some practical solutions', some researchers across the University may view Legal Services as a hindrance [$H_{16}$]. Some researchers appear to avoid seeking advice from legal advisors because: they seem to have a 'better to ask for forgiveness than to ask permission kind of mentality' [$J_{10}$]; and, they consider Legal Services to be a 'hindrance when they realise how much work is actually involved' to solve legal issues [$H_{16}$]. However, with or without the consultation of Legal Services any arising legal issues still need to be addressed by the researcher(s) [$H_{16}$]. **Miss J.** states:

> […] I make a point about being very well informed about [legal issues] these things, because […] I care about open science. […] I'm always more aware of when I need to contact a lawyer. But then I have to contact lawyers less often, because I'm frequently more clear about – what I'm doing – if that makes sense. Whereas, most scientists – they don't care, they want just to do their science. [$J_{10}$]

**Mr H.** states:

> […] I suppose the issue with them is the sheer-volume of contracts that we enter into. Contracts unfortunately aren't tailored – for each specific project as – as we would otherwise, I suppose, like to see them tailored. […] [$H_8$]

community norms will ensure that individuals are correctly attributed [$J_9$]. However, this leaves no avenue for recourse if someone chooses not to reference the data originator(s) [$J_9$]. Therefore, openly released data should not be confused with a data free-for-all; effective academic research data re-usage still requires some form of regulation.

#### 5.3.2.2 Multiple data originators

Crystallography is a highly collaborative discipline where the creation of multi-originated data is common.[367] Crystallographers are described by **Dr G.** as middle men [$G_{14}$], as **Dr A.** explains:

> […] [eCrystals is] being used by the National Crystallography Service [NCS] as a data repository. And so what happens is that users who've signed up to the NCS, or signed up to use the NCS, will send us crystals in, which we will then duly *er* run the data and solve the structure. And from then on, that structure then will go on eCrystals. […] [$A_{11}$]

The crystallographers that solve the crystal structures at the SCCG jointly own the results with the researchers that send the samples into the laboratory. In consequence, all these researchers have to reach a consensus on the terms and conditions of archival and if appropriate re-usage, as **Dr G.** describes:

> […] So they [other institutions] make stuff and we [the Southampton Chemical Crystallography Group] measure stuff *um* and you – you write a collaborative paper about making and measuring stuff. *Er*, so, we're completely reliant […] on the people who give us crystals […] and likewise. […] We do have a service level agreement […] on the national *um* facility that says: 'we'll keep it all forever *um* and after three years *er* keeping it private, we reserve the right to have a conversation with you *erm* about making it public'. *Er*, but it has to be on agreement from both parties. […] [$G_{14}$]

It is a key strength of eCrystals that all re-usage must be authorised by all collaborators directly involved with generating a crystallographic dataset. However, consensus is only one part of the difficulty. Collaborators must also be acknowledged for their levels of contribution [$A_{22}$, $G_{13}$]:

> […] People should be given *er* credit for any work they've done […] that goes without saying. And in many ways, […] we can have a problem with that in crystallography; in so far as that some chemists just see us as an outlet to a

[367] **Ms R.** highlights the difference between authorship practices within the humanities and the sciences:

> […] Humanities does [sic] mostly monographs. Whereas, computer science or *um* engineering and science disciplines, you would really, really look suspiciously at a paper with only one author – it would be a […] sign of a lack of quality usually, or of someone's personal interest […] we're going to see the same with data […] [$R_3$]

> technique, and so therefore don't want to actually give any credit for the work that's been done. *Um* but at the same time […] there does seem to be some papers with just massive, massive amounts of authors on – and yes, you wonder what they could've all done. […] [$A_{22}$]

Defining the roles of authors is therefore essential – and was a key problem in the case of Hwang and others.

#### 5.3.2.3 Embargos

From the primary source materials, the majority of data held by eCrystals is openly accessible. Only a small number of datasets are placed under short term or indefinite embargos. In the case of LabTrove, researchers can decide to release all their data openly or wait for their supervisor(s) to agree and sign-off its open release [$G_5$].

A robust embargo system is important to prevent the unauthorised release of academic research data before the researcher who collected the data has completed their initial project, and/or has published an article based on the data [$A_{13}$, $G_{13}$, $J_3$, $R_2$] (a point also highlighted by the MEDIN case study). Some researchers may use an embargo to prevent the release of data where they are concerned about its quality [$G_{14}$]. In the case of **Dr A.**, she embargoes all of her data on eCrystals 'until they are specifically published with reference to eCrystals' [$A_{13}$].

From the primary source materials, it is evident that academics are unable to analyse and publish papers about every single crystallographic dataset produced in the course of their research; there simply is not enough time. As a result some data will be embargoed for indefinite periods of time; for some datasets this is simply because these data have the potential to form part of a yet undetermined publication. eCrystals confronts this issue by permitting the automatic lifting on embargos [$G_{14}$], as the system was designed to enable long embargo periods which automatically make data open by default, unless the data originator keeps re-setting the embargo [$R_2$]. As **Ms R.** states: 'anything that was sort of forgotten would just be published by default rather than be left in the bottom of the metaphorical drawer' [$R_2$].

Some unauthorised releases of data can have serious legal and ethical ramifications, such as where a data leak could prove to be a contravention of data protection law or a trade secret [$G_{13}$, $G_{14}$, $J_3$]. Embargo systems therefore need to be risk-assessed for potential unauthorised releases of data. While only a minority of crystallographic data may form part of a patent application, for other scientific

disciplines the security of such data is extremely important on this basis, as **Miss J.** states:

> Patents will always have an impact. […] They'll have […] a two-fold impact from my understanding in that *erm* – you've got the first thing of well people don't really want that data releasing full stop – if they can have it. *Er* they're going to want it embargoed up until the date in which the patent is published at the very least – is the other effect. *Um* because of course, if we have an accidental release of data – something goes wrong – there's a malfunction in the system – whatever. *Er* then obviously that voids the patent, because information has been released in the public domain before the patent has been put in place. [$J_8$]

As eCrystals and LabTrove currently have a small remit and only manage crystallographic data that are generated via EPSRC funded research, these issues do not apply [$A_{15}$]. However, if the utilisation of eCrystals were to become more widespread, greater consideration would have to be given to the security and risk of managing data with vested commercial interests [$A_{15}$].[368]

**Ms R.** foresees the next open movement as open algorithms, therefore openly releasing the technical tools employed to collect the data [$R_2$], as is further explained in the following interview extract:

> […] You've got dataset A and now you've got dataset B – and that's all open. But the algorithm used to produce it you might want to patent, so you're not going to publish *um* – it might be commercially sensitive or, as happens in certain fields – it might be – the algorithm used was a commercial tool you were allowed to use, but you're not allowed to pass on. […] To take the same data and produce the same results, you need exactly the setup of libraries, the right copy of Windows and so forth. […] [In the past a colleague explored] archiving a virtual machine with the exact set-up. *Um* it – it would cost a lot to do, but in some – especially more controversial work, *er* where it might want to be proved later: 'well, here is the machine – we've bagged it up, we've saved it permanently. We don't have to keep the computer anymore, because we can save the computer as a virtual machine image – and it's just a big file with a hard drive and everything – and the exact copy of the process that you need to run it' […] [$R_2$]

It is clear that these algorithms would face similar issues to academic research data and publications, and therefore a robust embargo system would be required to protect any commercial and/or other interests.

---

[368] **Mr H.** contends that protection of sensitive and personal data types and commercial interests all have an impact on academic research data re-usage, such as confidentiality agreements between companies and the University where researchers re-use corporate data [$H_{19}$]. **Mr H.** insists where academic research data are offered on a commercial basis their quality needs to be assured [$H_{20}$].

#### 5.3.2.4 Research misconduct

There are no known examples of data misuse within eCrystals [$A_8$, $G_{10}$, $J_7$, $R_{11}$] or LabTrove [$G_{10}$, $J_7$]. Perhaps, this is because eCrystals is used by a small number of trusted individuals, and LabTrove is yet to mature. However, the potential for increased utilisation by unknown research individuals from more sites could change this [$A_8$]. eCrystals does not require research users to register their personal details [$G_3$] and (like MEDIN) the same is the case with LabTrove. Again as shown in the previous MEDIN case study, the enforcement of data licence agreements is problematic, as re-usage is reliant on goodwill and trust. Academic misconduct is therefore assumed to be handled at institutional level via procedures outlined by University policy. Legal action is a last resort – 'you don't rush on into litigation' [$H_{12}$].

Currently, the researchers inputting data into eCrystals are known and trusted individuals. In contrast, LabTrove is an open source piece of software which anyone can download and use to make all their research material visible [$G_{10}$]. It does not have a system in place to monitor re-usage [$G_{10}$]. As LabTrove is a blogging platform should it be treated as any other blog – as a neutral medium – which has a take-down policy for content where academic misconduct is proven [$G_{10}$]?[369]

#### 5.3.2.5 Attribution and licensing stacking

Although a general issue, **Mr C.** raised an important point about attribution and licensing stacking [$C_{11}$]. When analysing a vast amount of data from different sources, research users may decide to use text mining and/or data mining techniques to derive a new dataset. This new derived dataset and/or analysis may potentially have scores of contributing originators, attribution statements, and terms and conditions (some of which will be incompatible, for example assimilation of datasets that are and are not authorised for commercial re-usage purposes). **Mr C.** describes attribution stacking in greater detail, and explains why even though RCUK would prefer researchers to use one standard licence this will not necessarily resolve this attribution issue:

---

[369] **Mr C.** states:

> *Um* I suppose I see the Web as being *er* a kind of neutral space a bit like print would be […] the Web is just the medium. […] The QA part of it is all about the people interacting with the Web or the people interacting with the print. So you can't take the people out of the process. […] [$C_{25}$]

> […] the new RCUK policy, *um* on open access for publications, has specified the use of a CCBY licence. And, they're saying […] one of the reasons behind that is that they want to really maximise the re-use of those data, including *um* harvesting by text mining […] which I think is great in principle, but in practice you're still not getting 'round *er* the need for attribution. A lot of licences still say attribution is a requirement and […] I do think what you get […] attribution stacking. You get so many people that you need to attribute when you're doing this text mining – that in fact it becomes impossible very, very quickly to fulfil the attribution part of it. So you either need to look at a licence where you become, you know, not worried about the attribution – which a lot of people are concerned about, because you know then you're divorcing yourself from the kind of credit for the work as it were, and I fully you know understand that. *Um* but then if you hold onto the attribution as being key […] how do you get […] around this issue? So I think there's still a lot to do in that area. [$C_{11}$]

**Ms R.** further argues the problems with merging a large number of datasets:

> […] With research data […] as soon as you start merging data from multiple sources your data is subject to all of the restrictions of all of the sources. And is subject to all the requirements of all the sources, and *um* the more different licences, the more painful […] [$R_{15}$]

As with MEDIN, managing a number of different licences can be challenging, and while a one-size-fits all approach would be preferable this is hard to achieve [$H_{11}$]. This leads to a grey area of vital importance to re-usage in a digital age, and therefore requires greater research.

**Mr H.** also emphasises an important point about the problems not only with multi-party collaborations, but how cross-jurisdictional issues cause further complications as data are created across different jurisdictions with differing laws [$H_7$].

### 5.3.3 eCrystals and LabTrove's technological issues

#### 5.3.3.1 Functionality

From the interviews, it is evident that the technological framework underpinning eCrystals has a number of limitations.[370] eCrystals was selected as a test case for the application of EPrints to academic research data [$A_{18}$]. Therefore, the eCrystals archive runs using EPrints software, which was originally developed for the archiving of print publications rather than academic research data [$A_{18}$]. However, as the management and re-usage of crystal structures substantially differs from that of publications [$A_2$], eCrystals has a limited search functionality [$A_{12}$, $J_2$], usage and discoverability issues

[370] **Mr H.** emphasises that he is unclear about the exact systems and software issues – and it is therefore unlikely that Legal Services would be currently be aware of these also [$H_4$].

which may contribute to its limited number of users.[371] As an example of eCrystals' limited search functionality it does not enable research users to conduct sub-structure searches, which is a common feature in most other chemical inventory systems [$J_2$].

While there are plans to extend eCrystal's search functionality [$A_{14,}$ $A_{19}$]; previous plans to expand eCrystals failed to come to fruition [$A_{17}$]. However, it appears that a complete re-structure of its technological framework would be preferable to alteration of its current software [$A_{19}$, $J_{17}$]. The future development of its technological framework is dependent on funding [$A_{14,}$ $A_{19}$]. **Miss J.** states:

> […] Crystallographic data is very different to […] a print publication. So potentially I see it [eCrystals] possibly being completely re-designed […] as an eCrystals 2.0 thing – as a result of that there will probably be much more facility for metadata, […] sharing data. […] I predict that there will be much more fine grained permissions systems and *um* embargoing data will probably be made harder if anything. […] they'll probably be stricter licensing schemes in place […] [$J_{17}$]

Although there is a potentially larger audience for eCrystals outside the NCS and SCCG, the number of research users would be greater if eCrystals was more user-friendly for researchers depositing and re-using crystallographic data [$A_{20}$]. According to **Miss J.**, eCrystals would become as good as a data publication system subject to DOIs and a formal system of quality assurance [$J_6$].

While it is difficult to evaluate LabTrove's technological framework as it is in its infancy, the developers are aware that its interface requires improvement [$A_3$]. For instance, chemistry necessitates a significant amount of formatting; however, currently LabTrove only enables a basic text input [$A_3$]. For instance, since **Dr A**.'s involvement with the electronic laboratory notebook usability research project four years ago, she has only used her electronic laboratory once [$A_3$].

#### 5.3.3.2 Discoverability

From the primary source materials, it is clear that eCrystals is working to make its data more discoverable through the DataPool Project by testing the DataCite DOI Service, as **Mr C.** explains in further detail:

---

[371] However, there are still a number of technological improvements (such as infrastructure and software development) that need to be conducted across the entire University to facilitate greater academic re-usage data re-usage that not only meets funders' requirements but supports researchers in the course of their work [$C_4$].

> […] One of things we're [University of Southampton Library Services] doing with eCrystals is *er* they're helping us test the DataCite DOI Service with the British Library *er* with a view to *um* University of Southampton issuing digital object identifiers for data. *Um* because the National Crystallography Service that runs here at Southampton […] always has Southampton co-authorship and we're seen as the sort of custodians […] they're a – a prime candidate for testing the minting service. *Um* they would like to use it, they're very happy […] to work on *er* testing it, and then once they're happy with it we could then potentially roll it out across the University. We will of course have to develop criteria for when we will and won't […] assign a University of Southampton DOI […] to a data object […] [$C_1$]

Each dataset held within eCrystals and LabTrove has a DOI and a preferred citation that can be exported in a number of different formats (namely, oreChem, EndNote, BibTex and ASCII Citation) [$G_3$].

From the interviews, it is clear that DOIs are beneficial as they provide a way in which to establish a formal and persistent reference to datasets held within eCrystals and LabTrove [$C_2$, $G_3$, $J_6$, $R_{11}$]. This offers a user friendly approach, as individuals re-using the data have an easy way of providing an accurate and sustainable reference to a particular dataset upon re-use [$C_2$]. **Mr C.** and **Ms R.** contend that unique formal and persistent identifiers should not only be attached to data, but to individual researchers as well [$C_{13}$, $R_{11}$]:

> […] Some of the issues here are external too; it's not just about the institution, because the institution is not operating in a bubble. I think some improvements could be made nationally and internationally to the way things are done. So, you know simple improvements like having an – an identifier for researchers internationally. […] we don't even have, at the moment, a – a definitive list of funders that you could call off the systems […] [$C_{13}$]

Both **Mr C.** and **Ms R.** mention the non-profit organisation ORCID which provides 'a registry of unique researcher identifiers' [$C_{13}$, $R_{11}$].[372]

The reliance on one automatic system of referencing managed by a third party, such as DOIs, is only as effective as long as that system remains supported and sustainable however [$R_{11}$]. Therefore, is a re-direct system, such as DOI, sustainable to the same extent as using the Domain Name System (DNS) [$R_{10}$]? Or where the integrity of the ePrints URLs is maintained, is this the better system to use [$C_2$]? Both DOIs and

---

[372] 'About', *ORCID Website* <http://orcid.org/about> [accessed 9 August 2015], the website states:

> ORCID is an open, non-profit, community-based effort to provide a registry of unique researcher identifiers and a transparent method of linking research activities and outputs to these identifiers. ORCID is unique in its ability to reach across disciplines, research sectors, and national boundaries and its cooperation with other identifier systems.

DNS are distributed systems. Registration agencies are a significant feature of the DOI system; their function includes allocating prefixes and registering DOI names.[373] DataCite and CrossRef are two such DOI registration agencies employed within the academic community.[374] As with other identifiers (including DNS), DOI persistence is predicated on the person providing the content. DOIs could become invalid where the content is no longer the definition version; the International DOI Foundation (IDF) website offers the following guidance:

> **7. If I have assigned a DOI name and I make a change to my material, should I assign a new DOI?**
> The IDF does not have any rules on this. Individual RAs adopt appropriate rules for their community and application. As a general rule, if the change is substantial and/or it is necessary to identify both the original and the changed material, assign a new DOI name.[375] [IDF's bold emphasis.]

Broken links – especially link rot – are a problem for locating academic research data. While DOIs appear to be more stable, there is no one linkage system that is one hundred per cent reliable.[376] It remains crucial for people to assure the ongoing integrity of all links to their academic research data. Following on from this point, **Ms R.** also raises concerns about relying on third parties to manage the definitive source of data, in circumstances where the data originator has limited control over how those data are assured [$R_9$]. For instance, data may be hosted in another jurisdiction and therefore subject to different laws and agendas [$R_9$].[377] In this context, service level agreements are important as a formal agreement between data management service providers and

---

[373] 'DOI Handbook: 8 Registration Agencies', *International DOI Foundation Website* <http://www.doi.org/doi_handbook/8_Registration_Agencies.html> [accessed 9 August 2015]; 'DOI registration agencies', *International DOI Foundation Website* <http://www.doi.org/registration_agencies.html> [accessed 9 August 2015].

[374] 'History/mission', *CrossRef Website* <http://www.crossref.org/01company/02history.html> [accessed 5 July 2015]; DataCite Webite <https://www.datacite.org/node> [accessed 9 August 2015].

[375] 'Frequently asked questions about the DOI system', *International DOI Foundation Website* < http://www.doi.org/faq.html> [accessed 9 August 2015].

[376] For more information on persistent identifiers see: Emma Tonkin, 'Persistent Identifiers: Considering the Options', *Ariadne: Web Magazine for Information Professionals*, 30 July 2008 <http://www.ariadne.ac.uk/print/issue56/tonkin#15> [accessed 9 August 2015]; Laurence Horton, 'Digital Object Identifiers: Stability for citations and referencing, but not proxies for quality', *The Impact Blog: The London School of Economics and Political Science*, 23 April 2015 <http://blogs.lse.ac.uk/impactofsocialsciences/2015/04/23/digital-object-identifiers-stability-for-citations/> [accessed 9 August 2015].

[377] **Ms R.** further raises the point that if data are hosted in one jurisdiction alone – how stable is that country? For instance, war and/conflict could prevent research users from accessing data [$R_9$].

data originators (or their institutions where applicable) to set out agreed criteria for the management of specific datasets.

#### 5.3.3.3 Storage and delivery

eCrystals provides an adequate storage capacity for its existing number of datasets and the current pace at which additional datasets are generated. However, if the amount of data deposited began to increase dramatically, as is a known situation in the biomedical sciences for example [$C_4$], it is likely that its current storage capabilities would need to be re-evaluated.[378] Long-term preservation of data requires a considerable amount of staff time and effort in order to verify, annotate, curate and cleanse data [$C_3$]. Due to the requirement stipulated by University of Southampton's Data Policy that all significant data must be archived for a minimum of ten years, there are concerns over the necessary storage capacity and network requirements involved to uphold this policy [$H_{19}$]. This data archival situation is not unique to the University of Southampton, and is relevant across all higher education institutions.

Disciplines and institutions may have to work together to lower preservation costs and increase efficiency [$C_3$]; for instance, while the costs of storage may lessen over time it is likely that the cost of maintaining the data and metadata will remain [$R_3$]. While this is not currently an issue for eCrystals and LabTrove, where storage space is limited and/or there are not enough staff available to safeguard academic research data, decisions about what data are worthy of long-term preservation and length of data retention are extremely difficult and, to a certain extent, subjective. **Mr C.** explains how the potential long-term value of some datasets may be overlooked and ultimately hinder the advancement of knowledge:

> […] I think if something [an academic research datum] isn't used for twenty years, well it could be that in forty years' time, you know, it cures cancer and somebody wins the Nobel Prize. Just because something isn't used for twenty years doesn't mean that it is going to be a no use in the future. But equally you

[378] The biomedical sciences are already dealing with major storage issues, as a high amount of increasingly detailed imagery data is captured by researchers and therefore the size of each datum is rising exponentially [$C_4$]. Moreover, these large datasets are not only expensive, but technically challenging to store [$C_4$]. Therefore, it is not just the amount of research data and their rapid generation that is problematic, but the increasing size of some datasets. More detail can be stored and compressed, and new techniques enable the generation of richer data with more attributes than before. Therefore, academic research data re-usage models must remain cognisant of the storage challenges as hardware and software develops.

> have to make some decisions about the retention of your data, have policies in place for review of data […] [$C_3$]

Given the costs involved with data storage and preservation, attempting to accurately measure the (potential) value of an academic research dataset is imperative. There are many datasets where an application of a new technique or research methodology (e.g. the development of a new algorithm that is applied to a dataset) can derive new value even where in the short term the re-use of that dataset may be low. Hence quantitative metrics-driven models (e.g. how many times has a dataset been re-used in the past three months) should not be the only means of testing the worth of academic research data. There needs to be both a quantitative and qualitative assessment of its potential value. For that reason, value metrics is an important grey area for future research.

Following on from this point, the process of delivering large datasets can also be problematic [$C_{24}$]. While eCrystals facilitates instant delivery of data via downloads from the eCrystals website, MEDIN is not always able to provide instant downloads and in some cases still posts physical media (such as CDs and hard-drives) to research users. Therefore, while in theory the technological capability exists to transfer large files, the realities of expense and the lack of available technological means to deliver data instantly can hinder data sharing [$C_{24}$].

**Mr C.** and **Ms R.** both suggest how the use of BitTorrent to transfer large files could become standard academic practice to transfer large data files, but may not be popular partly due to its reputation for being used to move around unlawful content (such as copies of films and music that infringe copyright) [$C_{24}$, $R_{11}$]. **Ms R.** contends that data storage and delivery issues should be tailored to each individual case, as it depends on the general size and types of data being re-used [$R_{11}$].[379] The assumption that all openly accessible academic research data are digital and can be delivered to the research user instantaneously is a mistake [$R_3$].

---

[379] The physical storage and access requirements for academic research data varies across the disciplines [$R_3$]. For instance, a researcher within the health sciences may have accumulated a large number of personal data that are stored on a laptop and sealed in a safe until the end of the project where the hard-drive is put through a shredder and only the summative data are kept [$R_3$]. Another researcher within biomedical science may work with extremely large 3D high resolution images and therefore require temporary storage solutions as only the best images from a selection will be kept [$R_3$]. Whereas, the largest dataset another researcher manages may be a spreadsheet or posting data in real-time to a blog [$R_3$]. There is no one-size-fits-all solution [$R_3$] therefore academic models should examine the most efficient and cost effective ways to share data, for instance the average researcher could currently share a two terabyte file via a couriered hard-drive rather than a fibre optic cable [$R_{11}$].

There are also concerns over a potential data overload [$A_3$, $H_{14}$] (this issue is also raised in Chapters 2 and 4). **Dr A.** questions whether eCrystals stores too much data, and this richness obscures the discoverability of the most useful data [$A_3$]. **Dr A.** states:

> […] It was considered in the past really desirable to have everything possible – if you like, recorded on eCrystals – which *um* I think with hindsight *er* – for a crystal structure archive – you don't actually truly need […] everything. […] It gets to the point where finding the extra information is more time-consuming than the utility of the information. […] [$A_3$]

With regard to the Data Policy at the University of Southampton **Mr H.** considers: 'is it really useful to have so much information available [$H_{19}$]?' What data, metadata and level of detail are considered to be significant and worth archiving for future re-use is a substantial grey area. Is it enough just to publish the graph or should the underlying numbers be published as well [$R_{11}$]? Therefore, a decision for long-term data storage and preservation does not only just concern the quantitative and qualitative (potential) value of a dataset, but its previous versions, provenance metadata and other supporting materials.

### 5.3.4 eCrystals and LabTrove's socio-cultural issues

#### 5.3.4.1 Data sharing

From the primary source materials, it is evident that the majority of data generated through crystallographic experiments are not shared or made re-usable. Moreover, crystallography has endured tragic data loss as many crystal structures generated as part of shelved projects have never been published [$J_{13}$].[380] The culture of sharing data within chemistry is varied [$A_{23}$, $C_{22}$, $G_{16}$, $J_{20}$].[381] Researchers are more likely to make their data re-usable where it is a funding requirement, crucial to the publication process and/or enforced at a community level [$G_{16}$].

Within the sub-discipline of crystallography, there are a number of open data initiatives alongside eCrystals such as: (a) ChemSpider (a chemical structure database on the Web); and (b) CrystalEye (an aggregator of crystallographic data) [$A_{23}$, $C_{22}$, $J_{20}$]

[380] **Miss J.** asserts that the digitisation and deposit of these historic structures is a huge undertaking however – and perhaps, not the 'best use of a researcher's time' [$J_{13}$]. In general, data centres across the disciplines have lost funding in the past, putting academic research data into potential jeopardy [$C_3$].
[381] **Mr C.** believes that in principle a lot of academics support more openly accessible academic research data [$C_9$].

which has been incorporated into the Crystallography Open Database.[382] Since 1995, Chemical Mark-up language (CML) has been available as an open format [$J_{20}$].[383] However, while a significant proportion of chemistry research is collaborative, data sharing is often limited outside of a researcher's immediate project and network [$G_{16}$]. The interview participants proposed five main prohibitive (largely socio-cultural) factors for re-using data and reasons for not sharing data: (1) reward structure; (2) competitive advantage; (3) avoidance of error; (4) capture of non-conventional research materials; and (5) speed of releasing data to peers. These five prohibitive factors are now further outlined in the subsequent paragraphs.

As with many other disciplines, a key factor for taking a guarded [$G_9$] approach to sharing chemistry data stems largely from the reward structure [$R_2$]. Traditionally, higher education institutions and the Research Assessment Exercise (RAE – now known as the Research Excellence Framework (REF)) have placed greater emphasis on the

---

[382] *ChemSpider Website* <http://www.chemspider.com/> [accessed 5 July 2015]. The website provides a brief summary of ChemSpider: '*ChemSpider* is a free chemical structure database providing fast text and structure search access to over 34 million structures from hundreds of data sources.' *CrystalEye Website* <http://wwmm.ch.cam.ac.uk/crystaleye/> [correct on and last accessed 1 December 2013]. CrystalEye provides a brief summary of its role:

> The aim of the CrystalEye project is to aggregate crystallography from web resources, and to provide methods to easily browse, search, and to keep up to date with the latest published information. [CrystalEye Website].

Since the interviews, CrystalEye has been incorporated into the Crystallography Open Database – developed by Nick Day under the supervision of Peter Murray-Rust. *Crystallography Open Database Website* <http://www.crystallography.net/> [accessed 9 August 2015].

**Dr A.** states:

> *Um*, I think there's quite a big culture [of sharing] actually, on the whole. […] obviously there are certain aspects […] where data sharing doesn't go on […] perhaps amongst pharmaceuticals you know companies and that type of thing. Then there it's obviously a bit more closed, but certainly amongst academia I think there's […] quite a big culture of sharing. [$A_{23}$]

**Mr C.** explains:

> I think it [data sharing culture] depends on the part of chemistry […] if you're *er* working in crystallography *um* then you know there's […] not just […] eCrystals, but you've got ChemSpider […] and a number of initiatives […] about sort of open data within chemistry. But you also have parts of chemistry […] where *um* the prevailing *er* thought at the moment is that: 'we would like not to make this information open at all, because […] potentially we could be publishing off the back of this information for the next fifty years.' […] It's a very interesting discipline – you do have two completely kind of polarised views about where things are going. […] [$C_{22}$]

[383] **Miss J.** considers that to make academic research data re-usable they have to be stored in an open standardised data format with open source tools for parsing that standardised data format [$J_{12}$]. **Miss J.** believes that the majority of chemical data should be stored in some variant or extension of Chemical Mark-up Language which is an open source standard [$J_{12}$]. Refer to the following webpage for a number of useful articles by Peter Murray-Rust, Henry Rzepa and others: 'Chemical Markup Language – publications', *Chemical Mark-up Language – CML Website* <http://www.xml-cml.org/documentation/biblio.html> [accessed 9 August 2015].

value of research publications rather than underlying academic research data [$J_{24}$].[384] Therefore, researchers are aware that the impact of their research is judged on the quality of resulting publications, and not their academic research data [$R_2$]. Nevertheless, there is now greater emphasis placed on the value of underlying and supporting academic research data. The academic community has realised that academic research data can be released for re-usage after publication [$G_9$].

Chemistry researchers work within a very competitive academic environment and therefore many are cautious about sharing their data in order not to advantage their competitors [$G_9$, $J_{20}$].[385] There is a race to publish first and therefore researchers are often keen to safeguard their data and other findings [$A_{31}$].[386] For instance, if a researcher re-used a dataset within a highly-consulted publication, the researcher(s) who collected that data would receive minimal citation and impact factor [$R_2$]. This re-usage could be of potential detriment to their career [$R_2$].[387] Academic research data re-usage needs to enhance research impacts, and therefore researchers need to be made aware of its benefits not just on a personal or community level but for large-scale data analysis [$G_{19}$, $R_4$] (a point again raised by the MEDIN case study).

Some researchers are reticent to openly release their academic research data for re-usage, because they are sometimes cautious about data quality, such as inaccuracies and incomplete datasets [$C_9$, $R_5$]. However, as **Mr C.** explains in the following interview extract, academic research data are not perfect and errors will occur [$C_8$, $R_5$]:

> […] The best thing to do is to get into […] an area culturally where everyone recognises that nothing is perfect. So, there's a big difference between

---

[384] **Miss J.** states:

> Non-conventional publication […] next employer might not recognise your work […] under the next RAE exercise. [$J_{24}$]

[385] **Miss J.** states:

> […] I would say there is a pretty appalling culture of sharing within chemistry […] you can consider some things to be an invention – *um* and so there's a great deal of ownership that goes on […] This gets less the more computational you get. […] chemical experiments are quite easily globally repeatable […] therefore it would be very easy for someone to take your initial discovery and then scope your next three discoveries before you get there. […] And all this results in a mentality of protection, you don't want to share too much, you don't want to put your entire synthesis/method in your journal article, […] you don't want to dis-embargo your first crystal structure until your last crystal structure has been completed and published […] Having said that the Web is moving us forward […] the development of the CML format, […] ChemSpider and CrystalEye […] these things are making the data much more open. […] [$J_{20}$]

[386] **Dr A.** states:

> You can get gazumped […] [if] you're both working on something and then somebody else publishes before you. [$A_{31}$]

[387] **Dr A.** maintains that 'there's always been obviously a huge pressure to publish […] publishing mean prizes in that sense, whether it be jobs or whatever' [$A_{25}$].

> something being fraudulent *er* which clearly it's quite right that this kind of approach deters people from fraudulent practice, because they have to make a certain amount of their working available *um* to a community which is happy to correct each other's data […] another institution may spot that you've made a few mistakes in your data, but then – you know, the reverse will happen and in fact the benefit to the sector is that the data is much cleaner – and everybody has a much better quality based data to be working on. *Um* and that's the kind of collaborative approach to data cleaning which is very positive, but quite hard to achieve in a competitive environment, but I think […] it can be achieved. [$C_8$]

Cautious attitudes can constitute a prohibitive factor to sharing data. Researchers need to recognise potential for data enrichment that comes via community access and correction. However, to what extent this community-led approach to data cleansing is successful depends on the versioning and overall management of such datasets. Furthermore, researchers are more likely to keep their data in order where there is the possibility it will be re-used or released as open data [$R_2$].[388]

The primary source materials indicate that the traditional publication model is largely unable to publish research from all noteworthy crystallographic experiments. Not only are a small proportion of research findings published within crystallography, but around five to fifteen percent of findings from an experiment are published within a journal article [$G_{17}$].[389] Summarising this important point, **Dr G.** observes that traditional knowledge transfer within chemistry has been re-versioned as a digital process, but not re-purposed for the digital age:

> […] All we've done in chemistry is *er* turn the very conventional process electronic. So all we've done is *um* speeded things up, *er* made things more accessible in terms of the fact that *um* I can download it in seconds […] but effectively I'm getting the same level of information as a PDF as opposed to a photocopy. […] It's just an electronic version of what we've been doing for two hundred years. […]It [the Web] hasn't really impacted much on – on reliability, robustness, fitness for purpose, because we need access […] not to the PDF […] we need access to the actual data – *er* the actual observations, the actual recordings […] in the lab *er* to be able to *er* understand precisely what somebody did; and whether the […] inferences they made, the conclusions they drew, were correct or not. [$G_{16}$]

---

[388] **Ms R.** states:

> […] When you've got the assumption other people are going to see it […] from the start you tend to *um* keep it in slightly better order. […] If you're going to be judged by other people – you put that bit of effort in continuously. […] [$R_2$]

[389] **Dr G.** states: 'at the moment *er* what you do is only made available through the publishing process, and you can only publish a very small fraction of […] what you actually do – arguably somewhere between five and fifteen percent' [$G_{17}$].

Therefore upon access to a published crystallographic dataset, research users only obtain a condensed version of a research project or experiment. Core information framing the context of the experiment may however be missing, and will have a significant impact on the re-usability and repeatability of research. As **Dr G.** explains:

> [A publication will not] necessarily take into account, you know, what the temperature of the room was when you did a reaction for example […] and that can be crucial; you know, Helsinki in the winter versus Texas in the summer – could have a completely different outcome. [$G_{15}$]

Academic researchers may spend a year working on a research project, but largely have to distil all these data into a four page PDF to comply with academic journal stipulations [$G_{15}$]. Moreover, the structured ways in which crystallographic data are often published is a prohibitive factor for data re-usage [$J_2$]. For instance, in traditional chemistry journals – such as *Acta Crystallographica* – specific collections of crystal structures are published in each article (although *Acta Crystallographica Section E* often publishes individual crystal structures) [$J_2$]. Not having access to all the experiment data or data offered only in a pre-determined structure can, in some instances, stifle academic innovation.

As with many other disciplines, the speed at which academic research data are released to peers further hinders data re-usage. In the first instance, it often takes a long time between solving a crystal structure and its subsequent publication in an academic journal [$A_6$]. This slow release is often because it may take the duration of a PhD before a structure is submitted and/or researchers will only publish once they have compiled a group of structures to publish in one paper [$A_6$]. Furthermore, as journals often require potential research users to submit written requests for access to published crystallographic data, it can be very time-consuming to secure authorised re-usage [$J_{12}$]. **Miss J.** deems this journal data access process too slow and an unacceptable situation, as research users are unclear about the usefulness of a dataset until it is received [$J_{12}$].[390] In consequence, the publication process is certainly not taking advantage of the rapid knowledge transfer processes granted by the Web.

[390] The speed of release not only impacts on researchers within the sciences, but the social sciences and humanities. For instance, an academic within humanities may need to travel to an archive in another country to access historic data without knowing to what extent it may or may not enrich their research.

#### 5.3.4.2 Enhancing knowledge transfer

eCrystals and LabTrove are data capture and management systems [$G_{15}$] that aim to safeguard the complete series of data and metadata generated from crystallographic and other scientific experiments.[391] eCrystals and LabTrove embrace the concepts of open data, licensing, standards, protocols and software to form an open laboratory where research users are able to access the complete data series which is often either unpublished or partly missing from an academic journal article. The interviews not only raise the key prohibitive factors for re-using data, but proposed how eCrystals and LabTrove aim to confront these issues.

eCrystals and LabTrove can add value to research processes by enabling researchers from other laboratories not only to observe the types of experiments being carried out, but to compare the slightly different conditions and the effect this has on the outputs and findings [$G_{15}$]. **Dr G.** offers the following example to explain the benefits of greater data sharing:

> […] One PhD student might spend six months doing a reaction – but once you've got more information from other people who are trying the same thing, *um* ways in which you might be able to make each step quicker, more efficient – you could stream line the whole process – you could get a higher percentage yield faster – is what everyone is interested in. […] [$G_{15}$]

Therefore, the greater number of data available, the more potential there is for cross-comparison by different researchers and laboratories through data mining, use of statistical analysis and inference techniques across a wide-set of data [$G_{15}$]. This cross-comparison can be used to optimise and learn new experimental methods, to improve research efficiency and add further value to findings [$G_{15}$]. In twenty years' time, **Dr G.** would like to see the emergence of more open laboratories to foster widespread cross-

---

[391] **Miss J.** contends that academic research data has to be readily accessible by whatever medium, including the Web or other locations [$J_{12}$].
**Dr G.** states:

> […] [The] eCrystals system is basically a – a repository for instrument data. […] it captures *er* most of what's done in the laboratory […] and stores it. […] it's a mechanism for getting all the – all the stuff *er* that's gone on in the laboratory into one place – *er* and under *er* a single kind of structure. *Er*. It then has the ability to make those records more widely available […] So, the idea is that […] we capture and curate the data at source, as it's generated, […] as and when it becomes appropriate *er* we make it *er* public. […] [$G_2$]

**Dr G.** states:

> […] [The electronic laboratory notebook] captures everything you do every day *er* in the laboratory, and *er* as an individual – so it's a personal thing. *Er*, and then you can open it up […] to your supervisor, or line manager, […] immediate research group – *er* and you can make it available online as well. […] [$G_3$]

comparison [$G_{15}$]. However, it must be noted that while techniques such as data mining can advance academic understanding, they do not constitute a 'magic bullet' [$R_2$].

By helping research users to discover existing data [$A_{24}$], eCrystals and LabTrove have the further potential to help researchers form new alliances [$C_{23}$]. Researchers from institutions all over the world are able to access and assess a researcher's work and findings [$C_{23}$]. While data duplication is not a major issue for researchers within chemistry, it can be prevented through openly released academic research data [$A_{31}$]. For instance, in the past **Dr A.** has solved a crystal structure and subsequently realised that this had already been undertaken [$A_{31}$]. eCrystals and LabTrove are actively meeting EPSRC's requirements for greater access to data and demonstrate how the SCCG are 'maximizing the opportunity to do open access' [$G_9$].[392]

eCrystals and LabTrove offer alternative models of academic research data dissemination which, to a certain extent, circumvent the traditional academic publication process [$G_2$, $J_2$] through the provision of a relatively straight-forward route of publishing pure data findings without any implied conclusions (which belong within formal journal specifications) [$J_6$]. Tackling the issues previously raised where crystallographic data are often published as specific families by journals publishers, eCrystals publishes crystallographic datasets individually in their 'raw data format with the metadata' [$J_2$]. This enables the researchers 'to customise how you wish to think about families of crystal structures' [$J_2$], and is beneficial for collective analysis. This method is more conducive to current crystallographic research, which focuses less on the science of crystallography, and instead 'the science of structural systematics – why do things organise themselves as solid state matter as they do?' [$J_2$].

'Finding historical data in eCrystals is much easier' [$J_2$]. For example, **Miss J.** informs her PhD supervisor that she is examining a particular family of compounds [$J_2$].

---

[392] **Miss J.** states:

> […] You do need: the top-down mandating of *um* funding bodies, because it forces more reticent researchers to […] deal with open data; […] a strong community of researchers who […] do believe in open data and are willing to fight for it. […] the publishing systems we use *um* currently *er* are inherently biased towards a closed data format because they make their money from re-publishing other people's work […] [$J_{24}$]

**Mr C.** explains how mandatory open access is beneficial:

> […] [The Finch Report] did make a step in the right direction in terms of data, because it did say it would be looking, you know, also at *um* institutions making a statement about where their data are stored. *Um*, so that when you're publishing, you know, you'll be expected to say *um*, you know, what – what people could do to get hold of your data, and I think that's a reasonable thing to be able to do. If you've published a paper and there are data that support you know those assertions then it's not unreasonable […] to be able state where you've put those data. […] [$C_{14}$]

Her supervisor is able to support this research activity by pointing out ID numbers of datasets on eCrystals, which are part of that family [$J_2$]. **Miss J.** maintains that data on eCrystals can be shared between researchers quite readily through the use of Web links sent through email or social media [$J_2$].

eCrystals is different from other data archives, such as the Cambridge Structural Database, which for the past forty years [$G_2$] has trawled academic publications to extract and record crystallographic data [$A_7$, $G_2$, $J_{18}$].[393] While as far as **Dr A.** is aware they harvest eCrystals data [$A_7$], traditionally, structures on their database are finished with no more work will be done [$A_7$]. However, anyone can use eCrystals and build on that data [$A_7$]. **Dr A.** contends that a new question emerges – does the research user re-use a crystal structure on eCrystals or wait until a finished structure emerges – build and get a better solution [$A_7$]?

By offering open data repositories, eCrystals and LabTrove can be used to circumvent proprietary data providers including publishers and data aggregators that extract crystal structures from the published literature [$J_{20}$]. These data management systems can also further support academic publishers [$C_7$, $G_7$] by providing a 'comprehensive and rigorous form of electronic supplementary information' [$G_7$]. For instance, research users and aggregators do not have to manually or automatically extract data from publications [$G_7$]. In addition, academic journals do not have to repeatedly publish the same crystal structure(s) on each occasion there is a new observation [$J_6$]. Instead, academic publishers can provide links to datasets held in data repositories, such as eCrystals and LabTrove, within their journal articles via the normal referencing systems that are familiar to the scientific community [$J_6$]. Moreover, the researcher does not need to condense their entire findings into a four page PDF, as researchers can send the entire research process (all the data files generated throughout an experiment) [$G_7$].

From the interviews, it appears that there has been a perception among publishers that academic data re-usage platforms and other research repositories

---

[393] *The Cambridge Structural Database, The Cambridge Crystallographic Data Centre, University of Cambridge Website* <http://www.ccdc.cam.ac.uk/Solutions/CSDSystem/Pages/CSD.aspx> [accessed 9 August 2015]; 'Cambridge will always trawl the eCrystals for any data that they can, but they tend to only go for things which have formal publications with them' [$J_{18}$]. There is already a crystallographic database maintained by Cambridge, which obtains crystal structures from crystallographic research publications [$A_7$].

constitute a threat to their long-established model [$C_7$]. However, publishers seem to be increasingly open to data archives and repositories [$C_7$, $G_7$].[394] While it has been a long-standing role for the publishers to preserve the scholarly record, many seem concerned about over-committing to the safeguarding of underlying (and other supporting) data in perpetuity. Journal publishers may not be able to manage and maintain various formats and ultimately lack the knowledge to curate a dataset [$G_7$]. Safeguarding research publications is their primary concern.

As eCrystals and LabTrove both offer clear referencable records [$J_6$] through unique and persistent identifiers (DOIs), formal links to individual datasets held by these models are easily inserted into academic publications [$J_6$]. At the time of the interviews in 2012, **Dr G.** offers a further example of where a manuscript was sent to an academic journal publisher that contained links to all supporting data stored by an electronic laboratory notebook [$G_8$]. **Miss J.** is also aware of a published journal article where the authors make a number of references to the data held within their electronic laboratory notebook [$J_6$]. This is key advantage to further strengthen the robustness of peer-review processes in a digital age. In this example, the expert peer-reviewer was not only able to offer independent checks of the supporting and outlying academic research data, but provided valuable feedback to raise the quality of both the data and the paper [$G_8$]. This greater critique ultimately leads to better a publication which is of considerable benefit to academic publishers and the academic community alike [$G_8$]. Peer-reviewers do not normally have access to this level of underlying and supporting data and therefore are unable to make such comprehensive quality judgements [$G_8$]. This level of data scrutiny within the peer-review process is vital not just to rectify minor errors and omissions (e.g. by preventing the publication of data that has been poorly analysed [$G_8$]), but major incidents such as the case of Hwang and others.

While eCrystals and LabTrove are largely beneficial to the crystallographic and wider academic community, **Ms R.** nevertheless highlights potential concern for researchers using already existing corpora of data:

> […] a key part of science is repeatability […] actually having the entire steps to follow *um* – and just being able to see what other people did and compare every step not just the *er* final analysis – I think there's a real value in that. But at the same time *er* there's risks that I've heard put around by the community that if

[394] **Mr C.** makes the contention that there may be a different audience for data and publications, and therefore institutional and disciplinary repositories should not be considered as a threat to the traditional publication model [$C_7$].

> everyone uses the same corpus of data rather than generating their own data then you can end up with the danger of *um* an echo chamber built on top of incorrect information […] [$R_2$]

In consequence, research users cannot be fully reliant on the guarantee processes provided by the data originators, data managers and (where published within a journal article) peer-reviewers. It is of paramount importance that research users are responsible for personally scrutinising the data they wish to re-use. Open and transparent archives, such as eCrystals, offer opportunities for and encourage research users to inspect data structures by publishing extensive metadata and other materials.

#### 5.3.4.3 Continuity

In laboratory-based disciplines, project teams may encompass a principal investigator, research associates, laboratory technicians and/or PhD students. PhD involvement, grants and other commitments may result in a continual turnover of staff and/or more limited engagement with a research project from some parties. This potential disruption to personnel and project continuity can be problematic when trying to sustain a research project, and make it not just re-usable for others but usable for continuing and new staff.

Crystallography is a highly collaborative discipline and therefore protecting data for future re-usage in the event that the data originator leaves the project and/or organisation is highly important [$G_9$]. This is recognised by **Dr G.** as a key advantage of utilising eCrystals:[395]

> […] There's the re-usability by other members of the [eCrystals] team who generated it. […] Staff come and go – *er* and so other members of the team can pick up anything that's *er* has been done by somebody else before them […] the other aspect of re-usability *er* is making it available to the wider academic *er* community. Therefore, you know, anyone can come along and – and pick up these datasets *er* and do what they wish with them *er* as long as they acknowledge the source. […] [$G_2$]

The issue surrounding the continuity of personnel is not one that is specific to the eCrystals archive or an isolated problem for the research team a researcher leaves behind. Legal Services receive a number of queries from researchers who need to continue to use the data gathered during the course of their employment at a former institution [$H_3$]. Therefore, data models such as eCrystals and LabTrove facilitate

[395] **Mr H.** has received queries from researchers that need to continue using academic research data that was obtained partially or completely by another researcher who has since left the University and data were not deposited in an archive [$H_3$].

former researchers assessing data remotely where such data are openly accessible without the need for further rights clearance or special permission from their former institution.[396] Many of these issues can be resolved by prior planning, good record keeping, and rights management and clearance.

#### 5.3.4.4 Quality checks

The primary source materials indicate that eCrystals employs a traffic light system to signal different levels of data quality to research users: red alert – there are some serious problems with the data that require explanation; amber alert – there are potential problems with the data; and, green alert – there are potential areas for concern [$G_6$]. As crystallography is an experimental science, it is extremely rare to obtain perfect data with no quality alerts and so even the green alert carries a data warning [$G_6$]. Moreover, as it is also a very formulaic discipline, the final data generated are provided in a mark-up language that has 'a very particular, rigid format' [$G_6$]. This is advantageous because it can be automatically checked when uploaded to a free service on the Web to examine integrity and consistency of a crystallographic dataset [$G_6$]. This service offers an additional automated report, which alongside the data files supporting the report, are provided on eCrystals to offer a further level of data quality information [$G_6$].

As eCrystals has a small core user-base of around five researchers, most of whom are known to each other [$A_8$, $A_{16}$, $A_{34}$, $J_5$], the quality assurance procedures are more informal [$J_4$] and based on trust [$A_{12}$].[397] For instance, there is no procedure to verify that individuals have the authority to send and/or made the crystal structures delivered to the Southampton Laboratory for analysis [$A_{11}$].[398] Moreover, as LabTrove

---

[396] Researchers view academic research data as theirs, rather than the University's property [$H_{24}$]. For the most part, if a researcher leaves the University they will be able to use data they created during the course of the employment at the University of Southampton – especially where data are (agreed to be) made readily accessible subsequent to departure [$H_{24}$]. Legal Services have to ensure that there are no restrictions over this continued use, such as a contractual breach between the University and another organisation or issues of confidentiality [$H_{24}$].

[397] Therefore, eCrystals does not formally track the number of its research users [$A_{16}$]. **Dr G.** states:

> […] eCrystals now […] is used by the National Crystallography Service *er* and our […] research group here in Southampton. *Er* and in part, *er* a few other sites dotted around the globe […] but it's basically a – a laboratory repository. *Um* and the electronic lab notebook again – it came out of the same […] funding path. And, *er* that's now at the stage which we've got quite a few *er* groups around the world *er* using it […] [$G_1$]

[398] In general, **Mr H.** believes that research users should acknowledge that a dataset has been made in good faith and is provided on an as is basis [$H_{20}$]. The liability must rest with the user where the University has not been contacted directly and approved a specific re-use [$H_{20}$]. Readily accessible research data require a disclaimer for liability [$H_{21}$].

is relatively new it also has a limited number of users [$G_3$] where the level of quality assurance will be determined by personal settings.[399] From the interviews, it is clear that if eCrystals had a wider user-base it would require a stricter set of quality assurance criteria, as **Miss J.** explains:

> […] However, if we then extend [eCrystals] that out to people who are external – then – in a larger system you have to have formal checks and balances, because *urm* – you want to avoid – you know, not only genuine accidents but also to a certain extent you need to be able to avoid nepotism and – these other things which […] have undermined academic publishing historically. [$J_5$]

For both eCrystals and LabTrove data management is at the user level [$G_{18}$].[400] The individuals who are depositing their data in eCrystals are not only data originators, but data managers and data re-users [$A_4$].[401] The crystal structures that are deposited into eCrystals are independently reviewed by a higher qualified crystallographer within the research team [$J_4$]. As eCrystals runs using the EPrints software, it utilises the editor function, which means that on deposit a crystal structure is held in an editor's buffer until such a time that an independent crystallographer (within the research team) decides it is acceptable to openly release or embargo it [$G_5$]. However, it is the responsibility of the researcher depositing a crystal structure to ensure that it has been fully processed to a good quality and is scientifically sound [$A_4$, $J_4$].[402] Training new team members about

---

[399] **Dr G.** states:

> […] we have one or two groups, one of which is quite a high profile project, where they're basically doing open science on it. So everything […] from the lab goes straight into a lab notebook which is […] visible online by all at all times. And – and the funding for that project is *er* it's mandatory for everything to be completely upfront and open. [$G_3$]

[400] **Mr H.** maintains that it is not the responsibility of Research Innovation Services, the Library or Legal Services to provide quality assurance in terms of articles or data [$H_5$]. Their responsibility is to assure researchers and the University have the rights to deposit academic research data [$H_5$]. It is the responsibility of the researcher and their faculties to provide quality assurance policies and/or procedures [$H_5$]. **Mr H.** states there are no defined roles in the University for data managers [$H_5$].

[401] **Mr C.** states:

> […] You want it [the quality assurance process] to be rigorous in terms of risk, but you don't want it […] to be overbearing in that it becomes something additional to the researcher's work flow. *Um*, because ultimately your best QA is the people working within a university opt into the system readily because it's of help to them. And a QA system which is so rigorous that nobody uses it – and it's always bypassed – is no QA system; *um*, where people are just finding another way to do things because it's just too onerous to –to bother with a system that's been set up. […] [$C_6$]

[402] **Miss J.** outlines the responsibilities of data users:

> […] The responsibilities do obviously differ between data authors, data managers and data users. […] As a data user, you kind of have a responsibility also to *um* again do those same kinds of checks. Is this a scientifically sound piece of data I'm looking at? […] Because the temptation is just to disconnect – and say 'oh well this has been put in there, I know it has been peer-reviewed by higher crystallographers'. […] There are flaws – it's a human system at the end of the day, so mistakes can be made. So, you have a responsibility to double check on your data. And also to follow up on any re – references in formal journals and things as well to see if you know ... because

data quality and management further helps to avoid the release of poor quality data through eCrystals [$G_5$].[403]

As the majority of the data are openly released, these data are potentially open to greater scrutiny by research users [$A_{26}$].[404] Moreover, there is a statement on the homepage of the eCrystals website asking research users to report any errors within the data should they arise [$A_{30}$].[405]

## 5.4 eCrystals and LabTrove: interim conclusions

eCrystals and LabTrove aim to address the communal and continuity issues that arise through highly collaborative disciplines such as crystallography. eCrystals and LabTrove (depending on user preferences) provide a sustainable corpora of crystallographic data for re-usage by members of the SCCG and beyond. For instance, these data management systems protect data for future re-usage in the event that the data originator leaves the project and/or organisation. This continuity safeguard is vital for those working in laboratory environments where PhD involvement, grants and other commitments may result in a continual turnover of staff and/or more limited engagement with a research project from some parties.

A key strength of eCrystals is its consideration for all the parties involved with the generation of crystal structures, from those who collect the samples to those who process them. It ensures that all those involved are credited for their contribution, and are able to voice their opinions on the data management and re-usage parameters pertaining to a specific crystal structure.

---

if it's been published in a formal journal than obviously you've got *er* – that kind of *er* – a golden seal. [$J_4$]

[403] **Dr G.** states:

> […] Again the principle is that this is […] making data available by the people who do the science. So, the – the idea is that we're kind of self-regulating. *Urm*, nobody wants to release bad data out there that will *er* damage your scientific reputation. […] The control is really down to training people who are depositing stuff […] it's really about *er* personal responsibility more than anything. [$G_5$]

[404] **Dr A.** states:

> […] the fact that the […] Web has basically increased access to *um* to research, theoretically at least, should mean that it's actually more reliable and robust – because basically, obviously more people can check it. […] the flipside of that coin is because we can share so much data and we're swimming in a sea of data […] does anything get checked that closely anymore? […] in the past you didn't have anywhere near as much data – so consequently you could actually check all your data a lot more carefully […] [$A_{26}$]

[405] **Mr H.** hopes that research users would provide feedback in a positive way where data are not useful or accurate to enable researchers to address quality issues arising from their data [$H_5$]. **Mr H.** is unsure to what extent research users will offer feedback to the data owner over data quality [$H_{22}$].

In consequence, eCrystals and LabTrove overcome the major communal and continuity issues the occurred within the case of Hwang and others through robust attribution, licensing and data preservation safeguards. In addition, this case study highlights how such data management resources, that were lacking in the case of Hwang and others, are able to strengthen the journal peer-review process by: (1) providing the underlying and supporting datasets, provenance metadata and other processing information for independent scrutiny – and therefore facilitating the discovery of any potential bad or worst practices before publication; and, (2) raising the quality of a potential publication and its associated data via independent expert feedback.

While LabTrove proves a useful illustration of an alternative model to eCrystals, it is difficult to precisely evaluate its core frameworks, as it is set up in accordance with the personal preferences of the research user. As a result, while these interim conclusions evaluate both models, greater insight is available for eCrystals.

eCrystals has largely robust provenance metadata, legal and socio-cultural frameworks. For instance, each dataset held by eCrystals has a record of formulaic metadata and is openly released where possible under two familiar open licences, namely the Open Data Commons Attribution Licence and the Database Contents Licence. The crystallographic data generated are independently verified by members of the SCCG and published in open standards where research users do not need to learn or purchase new software. However, eCrystals has two areas that potentially require some further improvement: (1) its small user base, and (2) its IT infrastructure which was adapted from other usage. eCrystals runs via ePrints software, which was originally developed for the archival of print publications rather than academic research data. However, the research participants were extremely self-reflective about these issues, and considered options for future enhancements.

As with MEDIN, research users do not have to register to utilise eCrystals. A lack of resources means that there are no active checks to ensure that research users adhere to licence specifications. Again this reliance on goodwill and trust may prove to be a possible weak point of eCrystals, if in the near future its user base were to grow from the small group of current known users to a wider pool of unknown users.

From this chapter, it is clear that supply of provenance metadata is unable in itself to proactively prevent academic misconduct. Therefore, just because a model has a robust provenance metadata framework does not mean it can automatically block the release of erroneous data. Instead, provenance metadata should be considered as a

useful tool for retrospectively tracing back to sources of (possible) error. Moreover, this further highlights the potential for an echo chamber to be built around bad data (this was raised by **Ms R.**), because research users have confidence in a model 'as is' and therefore do not feel the need to fully scrutinise the provenance metadata provided.

As with the MEDIN case study, the same five themes emerge throughout the chapter – (1) sustainability, (2) discoverability, (3) working towards a common understanding, (4) accreditation, and (5) offering a good user experience – and are yet again raised in answer to the question: 'what makes for excellent quality academic research in a digital age?' These themes are now explained within this interim conclusion, and will be further evaluated as part of Chapter 7.

eCrystals and LabTrove were established to confront a number of sustainability and continuity issues by providing an open laboratory environment. For instance, eCrystals does not have to re-negotiate permissions with multiple data originators (whose contact details may have subsequently changed) for datasets that are not publically available where automatic embargo lifts have been arranged.

Those involved with eCrystals have also given significant consideration to the sustainability of links, as demonstrated by their involvement concerning DOI minting. Whilst there is no one solution to the problem of broken links, this case study has emphasised the need for people not only to think about the long-term future of their dataset, but the link providing access to it regardless of its type. For instance, persistent and unique identifiers can also be achieved through a higher education domain actively maintaining its URI integrity without reliance on third party persistent identification systems. In addition, this chapter underlines the potential value unique and persistent identifiers not only give to datasets, but researchers as well, such as through the ORCID initiative.

Although the search functionality provided by eCrystals is currently limited, discoverability remains an essential component of both eCrystals and LabTrove. eCrystals employs a system of machine-understandable metadata. Due to the formulaic structure of the crystallographic data, the provenance data are structured uniformly across each data record. In stark contrast, LabTrove allows research users to label their laboratory blog posts with user-generated tags, which organically join together with other posts in a many-to-many relationship. This chapter has been able to explore the contrast between the formulaic and structured provenance metadata generated by eCrystals, and the flexible and unstructured provenance metadata framework provided

by LabTrove. It appears – perhaps unsurprisingly – that the more unstructured provenance metadata becomes, the less discoverable its associated dataset also becomes. This is due to a lack of common understanding and standardisation, because research users using LabTrove may tag their data with synonyms. As a result, research users may only search for certain words and not their alternatives. In addition, it is more difficult to extract the scientific story from unstructured provenance metadata. However, the LabTrove developers are exploring technical solutions to help automatically derive more common understanding from the tagging system and ultimately improve data discoverability, without impeding on the flexible and personal character of an electronic laboratory notebook.

Within disciplines that are focused on data-driven experiments, such as crystallography, it is important for discoverability that provenance metadata can distinguish between several versions of the same dataset. For instance, a researcher may process the original dataset later on using different software and/or laboratory equipment. Moreover, provenance metadata needs to be able to manage intentional dataset duplicates where different laboratory groups may be working on identical datasets (often at similar times) for purposes of scientific validation and repetition. Therefore, for discoverability to be most effective, provenance metadata needs to be able to keep pace with the changing nature of definitive version(s) of datasets.

The eCrystals and LabTrove case study further highlights the importance of expertise-generated provenance layers to safeguard the quality and accuracy of academic research data. Consequently, provenance metadata are not just accounting for what a researcher has done. While it is vital for researchers to record crucial domain-specific provenance metadata, there are further required layers. Data managers may need to enhance institutional/bibliographic data with key words to attract a greater number of search queries. Legal experts may be required to sign-off the legal metadata to ensure there has been appropriate rights clearance and the dataset in question is released under an appropriate licence. However, with a lack of resources and time, such checks may prove unworkable for many data platforms. This raises the question: to what extent can closer links between stakeholders be forged? Is there a way in which an overarching system, providing a more seamless and integrated flow of information, could be created to produce an overview of the data process from collection through to supply and, in some cases, data destruction?

eCrystals and LabTrove are directly engaged with the open laboratory movement and are therefore committed to working towards a common understanding by using open standards and open source software. Moreover, there are a number of additional open data initiatives supported by the crystallography community, such as ChemSpider.

Great emphasis is placed on the potential misinterpretation of open licence systems, such as Creative Commons, during this chapter. While on the surface greater standardisation appears to be facilitating greater re-use, efforts are hampered by a lack of common understanding over legal issues. As with many higher education institutions, legal departments do not have the capacity to advise each researcher about the release and re-use of every dataset. Without this assistance researchers may confuse legal rights, or choose unsuitable licences. More work needs to be done to help raise legal awareness through greater education and on-demand guidance such as online case studies.

Accreditation is important to eCrystals as each dataset is independently and automatically checked through a free service on the Web that examines a crystallographic dataset's integrity and consistency. eCrystals also uses a traffic light system to further highlight the quality of the data and provides training to all joining colleagues on how to most appropriately use eCrystals. While it is difficult to fully understand whether accreditation is an essential feature of LabTrove, as a researcher can set up their electronic laboratory notebook with the purpose of releasing data only where it has been approved by a supervisor or colleague, accreditation is still an important aspect. However, it is recognised that if eCrystals were to attract a wider group of unknown depositors, existing best practice would have to be enhanced to accommodate members outside the current trusted network of five users, who are known to each other.

While currently eCrystals does not quite provide the level of utility desired, there is high level of appreciation for the extent in which the interview participants were self-reflective about its present limitations. Given this acknowledgement, it is clear that the provision of a good user experience remains important for both eCrystals, and LabTrove, and is central to its plans for improvement.

As well as the five key themes raised through the case study analysis, the interviews further highlighted three grey areas for further consideration: (1) meta-metadata – the potential for increased automated provenance metadata generation; (2)

unauthorised releases of data; and, (3) value metrics – the capability to estimate the potential future value of academic research data.

Firstly, it is likely that the next generation of laboratory equipment will increase the production of computer-generated provenance metadata. To ensure that these provenance metadata are reliable and fit for purpose would potentially necessitate some form of meta-metadata. This would be required in order to ensure that the machine is functioning properly and has attached the correct provenance record to the right dataset. How would models be able to manage such vast quantities of provenance metadata and provenance meta-metadata?

Secondly, the secure management of academic research data and provenance metadata are a concern for all re-usage models. Not only are there concerns over the extent to which provenance metadata and data can be de-linked in cases where datasets are copied and mined, but also over the unauthorised release of data. For personal and/or sensitive datasets, some unauthorised releases of data can have negative ramifications, such as where a data leak could prove to be a contravention of data protection law or a trade secret. Therefore, to what extent should academic research data models provide security measures and checks, and under which circumstances?

Lastly, this Chapter highlights problems over storage and delivery of academic research data and provenance metadata. Long-term preservation of data requires a considerable amount of staff time and effort in order to verify, annotate, curate and cleanse data. Therefore, if it is not possible to maintain all academic research data, how do researchers and their institutions determine the value of each dataset? It appears that with all the technological advances, important data are still at risk from data loss. A dataset that does not seem highly valuable at the time of collection could prove to be invaluable in the future when it is re-used and a new algorithm or piece of laboratory equipment is applied.

In summary, these data management systems – eCrystals and LabTrove – aim to re-purpose academic research data re-usage and their processes for a digital age by: (1) safeguarding and providing access to the experiment data that does and does not fit into the traditional four-page PDF journal article; and, (2) preserving the vast quantities of data that fall outside the traditional journal publication process for future re-usage i.e. where a researcher does not have time to produce a paper on every single crystal structure discovered. The thesis now turns to the final case study in Chapter 6 which

directly engages with ethical issues surrounding the re-usage of data gathered via human participants.

# Chapter 6: FLLOC and SPLLOC Case Study

The third significant area for attention concerning effective academic research data re-usage, raised in Chapter 2, is that of research ethics. In Hwang and others, unethical data were able to traverse the longstanding peer-review and other assurance processes that should have operated to halt its further dissemination. While it is evident in the MEDIN, eCrystals and LabTrove case studies that breaches of research ethics had never been an issue nor a major concern, this final case study focuses on such questions.

This chapter provides an insight into data generation (via human participants) within modern languages research that requires robust ethical processes, as this research involves minors and other sensitive situations. Academic research data gathered from human participants are amongst the most difficult to release or fully disclose as open data; especially where those data are personal and in some cases extremely sensitive. In particular circumstances data sharing can impact on personal safety, for instance where participants are political dissidents or whistle-blowers. Therefore, this chapter now focuses on: how the maintenance of and access to sensitive and personalised data should be treated in order to balance a range of difficult issues with permissions, data protection and confidentiality.

The French Learner Language Oral Corpora (FLLOC) and the Spanish Learner Language Oral Corpora (SPLLOC) are two data platforms for the re-usage of second language acquisition data, and are the central focus for this chapter. Second language acquisition research falls largely under the discipline of modern languages, and focuses on the ways in which young learners acquire non-native languages.[406] FLLOC and SPLLOC provide access to second language acquisition research data collected largely from human participants under the age of sixteen that are available to anyone with a Web connection. It is stated on FLLOC website that:

> The contents of the database are being made freely available to the research community, in the form of digital sound files and related transcripts […] The database currently contains over 4000 files (sound files, transcripts and morphosyntactically tagged transcripts).[407]

---

[406] *French Learner Language Oral Corpora (FLLOC) Website* <http://www.flloc.soton.ac.uk/> [accessed 9 August 2015]; *Spanish Learner Language Oral Corpora (SPLLOC) Website* <http://www.splloc.soton.ac.uk/> [accessed 9 August 2015].

[407] *FLLOC Website*.

This is reiterated on the SPLLOC website:

> The corpus has been designed in a balanced way to fill the existing methodological gap in Spanish L2 resources. […] The dataset of audio files and accompanying transcriptions in CHAT format created through the SPLLOC 1 project is now publicly available.[408]

By capturing learners through video/sound recordings and transcription, researchers are able to critically assess real-world examples of second language acquisition. FLLOC focuses on British individuals learning French, whereas SPLLOC concentrates on British individuals learning Spanish. This research is not only of benefit to academics within modern languages, but education professionals such as teachers, lecturers and policy makers.

Comprising nine research projects that have taken place over a twenty-four year period, FLLOC is undoubtedly an open data pioneer. Established in 2006, SPLLOC is FLLOC's sister website that is composed of two research projects. FLLOC and SPLLOC focus on the collection of data from human participants, therefore it is clear that this research must remain compliant with the Data Protection Act 1998, and the requirements, frameworks and codes enforced by the higher education institutions involved. As many of the these researchers were employed by the University of Southampton, Chapter 3: Methodology provides a comprehensive overview of research centred on the collection of personal data fulfilling the requirements prescribed by the University of Southampton's ethical procedures.[409]

While this case study focuses on the overarching provenance metadata, legal, technological and socio-cultural frameworks employed by FLLOC and SPLLOC, it further examines the two most recent FLLOC projects and corpora: (1) the Young Learners Corpus, and (2) the LANGSNAP Corpus. The Young Learners Corpus

---

[408] 'Rationale', *Spanish Learner Language Oral Corpora (SPLLOC) Website* <http://www.splloc.soton.ac.uk/rationale.html> [accessed 9 August 2015].

[409] While the vigorous ethics procedures remain unchanged, since this thesis was approved through the paper system a new digital platform called Ethics and Research Governance Online (ERGO) became the new 'electronic document handling system for Ethics forms, IRGA forms and any other supporting documentation'. *Ethics and Research Governance Online (ERGO), University of Southampton Website* <https://www.ergo.soton.ac.uk/> [accessed 9 August 2015]. For more information about the University of Southampton's ethics procedures refer to: 'Research Governance Office (RGO)', *University of Southampton Website* <http://www.southampton.ac.uk/corporateservices/rgo/> [accessed 9 August 2015]. For information about the Newcastle University's ethics procedures refer to: 'Research and Enterprise Services: Ethics in the University', *The Newcastle University Website* <http://www.ncl.ac.uk/res/research/ethics_governance/ethics/procedures/university/ethics_university.htm> [accessed 9 August 2015].

collected data from British schoolchildren, five to eleven years old, learning French (the SPLLOC 1 project is very similar). By contrast the LANGSNAP Corpus gathered data from undergraduate modern languages students on their intercalated year abroad as part of their BA modern languages degree course at a British university. The LANGSNAP Corpus centres on the impact of social networks on young adult learners (mainly in their early twenties) acquiring non-native languages. As these two projects are the most recent, they enable access to participants with both current insight and direct connections to the historical development of FLLOC and SPLLOC. Moreover, Professor Rosamond Mitchell and Professor Florence Myles – who are credited with creating FLLOC and are the co-directors of SPLLOC – both led the Young Learners research team, and Professor Mitchell was the principal investigator of the LANGSNAP project. Therefore, the two projects have a level of continuity through the co-directors' involvement, and were both subject to the ethics procedures at the University of Southampton.

By first examining the primary source materials, this case study investigates the overarching provenance metadata as well as the legal, technological and socio-cultural frameworks determined by FLLOC and SPLLOC as indicative of further project-specific frameworks, which are tailored to the specific context of each corpus (managing minors and young adults as participants). This critical analysis highlights the key issues that require further clarification by the six interview participants.

FLLOC and SPLLOC involved the collection of personalised and not necessarily sensitive data, as defined by the Data Protection Act 1998, section 2 which states:

> In this Act "sensitive personal data" means personal data consisting of information as to— [/] (a) the racial or ethnic origin of the data subject, [/] (b) his political opinions, [/] (c) his religious beliefs or other beliefs of a similar nature, [/] (d) whether he is a member of a trade union (within the meaning of the M1Trade Union and Labour Relations (Consolidation) Act 1992), [/] (e) his physical or mental health or condition, [/] (f) his sexual life, [/] (g) the commission or alleged commission by him of any offence, or [/] (h) any proceedings for any offence committed or alleged to have been committed by him, the disposal of such proceedings or the sentence of any court in such proceedings.[410]

[410] The Data Protection Act (DPA) 1998, section 2. *UK Government Legislation Website* <http://www.legislation.gov.uk/ukpga/1998/29/section/2> [accessed 9 August 2015].

This case study highlights further issues faced by modern languages academics (and other academics across the sciences, social sciences and the humanities) gathering data of a sensitive nature from human participants. The research area of life histories provides an example of where these issues are confronted. Life histories offers a way in which researchers can access, record, analyse and interpret the experiences of particular individuals. Life histories research is wide-ranging from capturing memories of the Cuban Revolution to individuals' memories of the University of Warwick from 1965-2015.[411]

Alongside personalised data, life histories researchers sometimes encounter extremely sensitive data gathered from adult and child participants through interviews, memoirs and other personal materials. Such participants may be political dissidents or whistle-blowers, where to uncover their true identities could have a major impact on their personal safety and that of their families, their employment, and/or their reputation. It is of paramount importance therefore that life histories researchers protect the identities of their participants, and remain compliant with the Data Protection Act 1998. Moreover, participants may express socially objectionable, intolerant and prejudiced views – such as sexist, racist and homophobic opinions – that require particular ethical handling procedures. As a result, there are considerable ethical responsibilities tied up with this type of personal sensitive data collection and the research investigating it.

---

[411] For further background information about life histories research refer to: 'Centre for Life History and Life Writing Research', *University of Sussex Website* <http://www.sussex.ac.uk/clhlwr/> [accessed 9 August 2015]; for some examples of life histories research refer to: 'National Life Stories', *British Library Website* <http://www.bl.uk/nls> [accessed 9 August 2015]; 'Memories of the Cuban Revolution', *University of Southampton Website* <http://www.southampton.ac.uk/cuban-oral-history/> [accessed 9 August 2015]; 'Voices of the University: Memories of Warwick, 1965-2015', *The University of Warwick Website* <http://www2.warwick.ac.uk/fac/cross_fac/ias/current/universityvoices/> [accessed 9 August 2015]; 'Scottish Oral History Centre', *University of Strathclyde Website* <http://www.strath.ac.uk/humanities/research/history/sohc/> [accessed 9 August 2015]; 'African Oral History: Oral history across generations – A research programme with the universities of Dakar and Algiers', *University of Portsmouth Website* <http://www.port.ac.uk/research/africanoralhistory/> [accessed 9 August 2015]; 'Sound collection', *The Imperial War Museum Website* <http://www.iwm.org.uk/collections-research/about/sound> [accessed 9 August 2015]; 'Nurses' lives: the oral history of nurses', *Kingston University, Faculty of Health and Social Care Website* <http://www.healthcare.ac.uk/research/nurses-lives/> [accessed 9 August 2015].

**Dr K.** who is an interdisciplinary academic at the University of Southampton involved with life histories research was selected to accentuate the more sensitive nature of political data collection within modern languages research, and therefore further enhance this case study.[412]

This case study has three strands; it first evaluates the safeguards in place that facilitate the re-usage of data from vulnerable participants – children – who are unable to personally give their consent (under the age of sixteen, as it is the parent or guardian's right to give or withhold consent on their child's behalf). Secondly, it focuses on young adults who are coping with living abroad (in many cases) for the first time, and are therefore (potentially) experiencing new challenges both socially and through their studies. Finally, the addition of life histories research captures a type of data that are often personal and extremely sensitive. Through critical assessment of the primary source materials and the interviews, this case study considers to what extent these case study safeguards rectify the last major issue raised by the Hwang case, and how this model could resolve potential issues which manifest in current and future cases of bad and worst practice.

[412] It must be noted that personal sensitive data are not only collected through humanities, but by researchers across the sciences and social sciences.

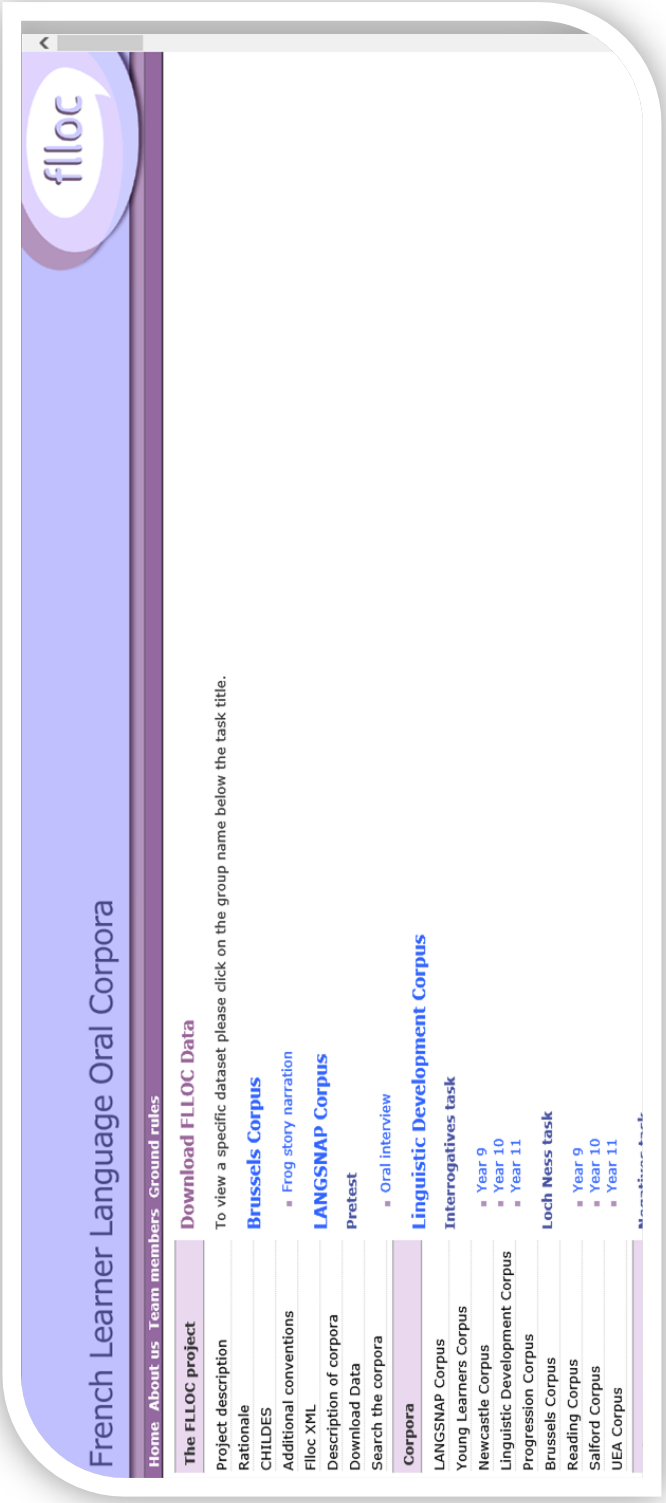



**Figure 4 Screen shot of 'Download FLLOC data', *FLLOC Website* <http://www.flloc.soton.ac.uk/tasklist.html> [accessed 1 November 2014].**

**Citation: Professor Florence Myles and Professor Rosamond Mitchell and the other corpora owners <http://www.flloc.soton.ac.uk/> and the 'CHILDES Project' <http://childes.psy.cmu.edu/> [accessed 1 November 2014]. Image taken on 1 February 2014. Reproduced with permission from FLLOC, University of Southampton.**

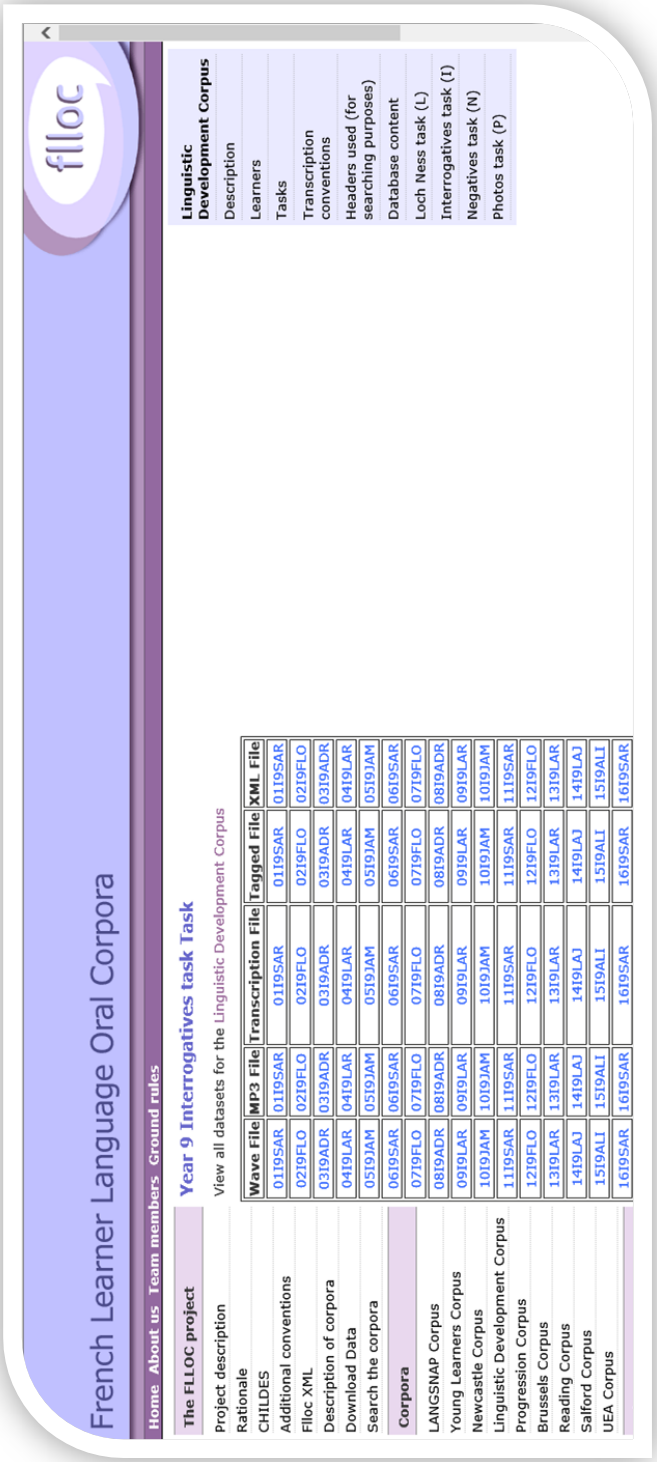



**Figure 5 Screen Shot of 'Year 9 Interrogatives Task', *FLLOC Website* <http://www.flloc.soton.ac.uk/ldc/datasets/LDCI9.html> [accessed 1 November 2014].**

**Citation: Professor Florence Myles and Professor Rosamond Mitchell and the other corpora owners <http://www.flloc.soton.ac.uk/> and the 'CHILDES Project' <http://childes.psy.cmu.edu/> [accessed 1 November 2014]. Image taken on 1 February 2014. Reproduced with permission from FLLOC, University of Southampton.**

## 6.1 FLLOC and SPLLOC: primary source materials

### 6.1.1 FLLOC and SPLLOC's rationale

FLLOC and SPLLOC have a number of readily accessible second language acquisition data available on their websites in the form of transcripts, sound files and task outlines from eleven projects (an overview is given below). They are two sister websites hosted by the University of Southampton, and overseen by Professor Rosamond Mitchell (University of Southampton) and Professor Florence Myles (University of Essex, formerly University of Southampton and University of Newcastle) who are both academics in the field of second language acquisition research.[413]

While the first FLLOC project – Progression in Foreign Language Learning – collected data from British schoolchildren learning French from 1993-6, the FLLOC website was not created until – the Linguistic Development in Classroom Learners of French: a Cross-Sectional Study – which took place during 2001-2. However, the sound recordings from the original project are now openly available on the FLLOC website under the Progression Corpus.

FLLOC brings together nine individual research projects. The following five corpora are directly linked to Professor Mitchell and Professor Myles, and the Universities of Southampton and Newcastle: the LANGSNAP Corpus (2011-3); Young Learners Corpus (2009-2011), Newcastle Corpus (2005-8), Linguistic Development Corpus (2001-2); and the Progression Corpus(1993-6).[414] Four related projects were conducted by researchers at other universities: Brussels Corpus (1994-5), Reading Corpus, Salford Corpus (1989-1993); and the UEA Corpus (2002-4).

SPLLOC consists of two projects: SPLLOC 1 (2006-8); and SPLLOC 2 (2008-2010). These collaborative projects were conducted by second language acquisition researchers from the University of Newcastle, the University of Southampton and the University of York. SPLLOC therefore connects to the original aim of FLLOC by conducting similar research but within the domain of non-native learners of Spanish.

---

[413] 'Modern Languages: Our Staff – Professor Rosamond Mitchell', *University of Southampton Website* <https://www.southampton.ac.uk/ml/about/staff/rfm3.page?> [accessed 9 August 2015]; 'Department of Language and Linguistics: Academic Staff – Professor Florence Myles', *University of Essex Website* <http://www.essex.ac.uk/langling/staff/profile.aspx?ID=2332> [accessed 9 August 2015].

[414] Dr Laura Dominguez was also the principal investigator on the 'SPLLOC 1 Project', *FLLOC Website.*

The principal rationale behind FLLOC and SPLLOC is to build a re-usable resource of quality, digital, second language acquisition research data that can be shared with the research community (and other interested parties) in the long term. In the words used by the MEDIN interviewees, FLLOC and SPLLOC are centred on gathering data once, and re-using many times.

### 6.1.2 Young Learners Corpus: overview

From September 2009 to September 2011, the Economic and Social Research Council (ESRC) funded a further project: Learning French from ages 5, 7 and 11: An investigation into starting ages, rates and routes of learning amongst early foreign language learners.[415] This research project led to the formation of the Young Learners Corpus, and involved researchers from both the University of Newcastle and the University of Southampton.[416] Professor Rosamond Mitchell and Professor Florence Myles led the research team as co-directors (both academics were originally colleagues at the University of Southampton).[417] The project had four key objectives:

> Document the development of linguistic competence among young classroom learners of French at three different starting ages, in primary and early secondary school classrooms, and identify similarities and differences. [/] Compare rates of development at different ages after the same amount of classroom exposure [/] Document and compare the classroom learning strategies used by children at different ages and their attitudes to language learning [/] Through this evidence, contribute to theoretical understandings of second language acquisition among

---

[415] Florence Myles and others, 'Learning French from ages 5, 7, and 11: An Investigation into Starting Ages, Rates and Routes of Learning Amongst Early Foreign Language Learners', ESRC End of Award Report, RES-062-23-1545 (Swindon: 'Economic and Social Research Council' (ESRC), 2012), *The Economic and Social Research Council (ESRC) Website* <http://www.esrc.ac.uk/my-esrc/grants/RES-062-23-1545/outputs/read/01559a37-eb0c-409a-afdf-28d515758a88> [accessed 9 August 2015]; 'Impact and Findings: Research Catalogue – Learning French from ages 5, 7 and 11: An investigation into starting ages, rates and routes of learning amongst early foreign language learners'. RES-062-23-1545 *The Economic and Social Research Council (ESRC) Website* <http://www.esrc.ac.uk/my-esrc/grants/RES-062-23-1545/read> [accessed 9 August 2015]; *FLLOC Website*; 'Young Learners Corpus: Description', *French Learner Language Oral Corpora (FLLOC) Website* <http://www.flloc.soton.ac.uk/primary/index.html> [accessed 9 August 2015].

[416] *The Economic and Social Research Council' (ESRC) Website* <http://www.esrc.ac.uk/my-esrc/grants/RES-062-23-1545/read> [accessed 9 August 2015].

[417] 'Young Learners Corpus: Description', *French Learner Language Oral Corpora (FLLOC) Website* <http://www.flloc.soton.ac.uk/primary/index.html> [accessed 9 August 2015].

> young learners, and consequently inform current primary language initiatives and educational practices in the UK and internationally.[418]

The Young Learners Project collected second language acquisition research data from seventy-two pupils learning French (for the first time) in two schools in the North East of England: twenty-seven of the schoolchildren were five years old (Year One); twenty-six of the schoolchildren were seven years old (Year Three); and, nineteen of the schoolchildren were eleven years old (Year Seven).[419] In consequence, it was crucial for the researchers to obtain consent from the parents and schools. Moreover, the researchers required Criminal Record Bureau (CRB) checks as they were working with minors.[420]

The researchers 'provided 38 hours of French language teaching for each of those three groups of learners'.[421] The children were recorded undertaking a number of tasks designed by the research team. Data are available for the elicitated imitation, story-retelling and role play activities through FLLOC (for detailed descriptions of these activities see footnote below).[422]

On the FLLOC website, data from these tasks are organised by year groups, type of tasks and the four different stages of collection. For all the tasks, the sound recordings are available as wave and MP3 files. The elicitated imitation task data have additional tagged (textual) files available for download. The role play and story-retelling activities data have transcript (textual) files for download also, and there is an option to download zip files containing these multiple datasets.

### 6.1.3 LANGSNAP Corpus: overview

From May 2011 to October 2013, a research project called: 'Social networks, target language interaction and second language acquisition during the year abroad: a

---

[418] 'Young Learners Corpus: Description'.
[419] 'Young Learners Corpus: Learners', *French Learner Language Oral Corpora (FLLOC) Website* <http://www.flloc.soton.ac.uk/primary/learners.html> [accessed 9 August 2015].
[420] CRB checks are now known as Disclosure and Barring Service (DBS) checks.
[421] 'Young Learners Corpus: Description', *French Learner Language Oral Corpora (FLLOC).*
[422] 'Young Learners Corpus: Tasks', *French Learner Language Oral Corpora (FLLOC) Website* <http://www.flloc.soton.ac.uk/primary/tasks.html> [accessed 9 August 2015]. The elicitated imitation activity is where a learner is given a picture, played a set of recorded sentences individually in French, asked to repeat the sentence heard, and finally asked two comprehension questions. The story re-telling activity involves learners re-telling a specific story in French, previously practised as a class, with pictorial prompts. The role play activity requires learners to work as part of a pair or small group with a researcher. For instance, the researcher uses a doll or cartoon character (gender and age appropriate) as a prop, which is used to ask the learners 'about their names, ages, families and hobbies' in French.

longitudinal study' was funded by an ESRC.[423] This project led to the formation of the LANGSNAP Corpus, and involved researchers from the University of Southampton. Professor Rosamond Mitchell led the research team as the director.[424]

> The specific aims of the LANGSNAP project were to document the development of advanced level students' knowledge and use of L2 French or L2 Spanish over a 21-month period including a 9-month stay abroad, and to make the resulting learner corpora available freely to the international research community.[425]

The LANGSNAP project collected data from fifty-six modern languages undergraduates learning French or Spanish at a British University whilst they spent a year abroad (as part of their degree programme) in France, Spain or Mexico.[426] Ten native French and Spanish speakers also completed the same language tasks for the purposes of comparison.[427] These participants were recruited through ERASMUS (the European educational exchange programme), foreign language assistant positions, and other work placements.[428] The LANGSNAP researchers collected data through interviews, language tasks and questionnaires at six stages throughout the lifecycle of the participants' time abroad. [429]

In addition, the researchers also selected twelve of these undergraduate modern languages students to participate in individual case studies which involved researchers observing the participants on two separate occasions during their year abroad, and the participants recording themselves in discussion with their peer group in French/Spanish on three separate occasions with provided audio recording equipment.

The LANGSNAP corpus was not available at the time of the interviews. This is because the project had yet to complete its final report when the thesis' interviews were conducted. The willingness of the FLLOC and SPLLOC interview participants to share

---

[423] 'Social networks, target language interaction and second language acquisition during the year abroad: a longitudinal study', ES/l018298/1 or RES-062-23-2996, *Research Councils UK Website* <http://gtr.rcuk.ac.uk/project/5F8473E8-E1CE-4531-AE78-72719ECD1BA5> [accessed 5 July 2015]; 'Description: LANGSNAP Corpus', *French Learner Language Oral Corpora (FLLOC) Website* <http://www.flloc.soton.ac.uk/LANGSNAP/index.html> [accessed 9 August 2015].

[424] 'Description: LANGSNAP Corpus', *FLLOC Website.*

[425] 'The LANGSNAP Project', *LANGSNAP Website* <http://langsnap.soton.ac.uk/project.html> [accessed 9 August 2015].

[426] 'The LANGSNAP Project', *LANGSNAP Website.*

[427] *Ibid.*

[428] *Erasmus+ Website* <https://erasmusplus.org.uk/llp-and-youth-in-action/erasmus> [accessed 9 August 2015].

[429] 'The LANGSNAP Project', *LANGSNAP Website.*

their insights prior to publication of results, further reveals that data gathering is as valuable as data analysis ensuing from it (a point emphasised by Chapter 5: eCrystals and LabTrove Case Study). The LANGSNAP data is now available through its website; users can browse the data based on the activity, collection round and participant.[430]

### 6.1.4 CHILDES: overview

The historical development of second language acquisition research (as is the case with many other disciplines) is deeply entwined with technological innovation. Before the materialisation of the tape recorder in the 1950s, 'even the most highly trained observer could not keep pace with the rapid flow of normal speech production'.[431] While the tape recorder enabled the researcher to capture a complete oral dataset, this did not solve the problem of data dissemination.[432] In order to share the oral data, researchers had to transcribe the sound recordings via written and/or typewritten copies.[433] As was the case in MEDIN, this led to a lack of standardisation, as different researchers used their own transcription coding systems.[434]

As a result of the greater accessibility to and utilisation of computers, the Child Language Data Exchange System (CHILDES) – founded in 1981 and launched in 1984 – was developed to resolve these data sharing issues:

> In 1983, the MacArthur Foundation funded meetings of developmental researchers in which Elizabeth Bates, Brian MacWhinney, Catherine Snow, and other child language researchers discussed the possibility of soliciting MacArthur funds to support a data exchange system. In January of 1984, the MacArthur Foundation awarded a two-year grant to Brian MacWhinney and Catherine Snow for the establishment of the Child Language Data Exchange System.[435]

---

[430] 'Browse data', *LANGSNAP Website* <http://langsnap.soton.ac.uk/view/> [accessed 9 August 2015].

[431] Brian MacWhinney, 'The CHILDES Project: Tools for Analyzing Talk – Part 1: The CHAT Transcription Format', *CHILDES Manual*, Carnegie Mellon University (19 June 2015), p. 6 <http://childes.talkbank.org/manuals/CHAT.pdf> [accessed 9 August 2015].

[432] *Ibid*, p. 7.

[433] *Ibid.*

[434] *Ibid.*

[435] 'The CHILDES Project: Tools for Analyzing Talk – Part 1: The CHAT Transcription Format', p. 8.

FLLOC and SPLLOC are subscribers to CHILDES, which comprises three main components: Codes for the Human Analysis of Transcripts (CHAT); Computerized Language Analysis (CLAN); and, the CHILDES database.[436]

CHILDES is one of the international TalkBank databases that contain data from human and animal communication, which are coordinated by Professor Brian MacWhinney at Carnegie Mellon University in the USA.[437] TalkBank is described as

> an interdisciplinary research project funded from 1999 to 2004 by a grant from the National Science Foundation (BCS-998009, KDI, SBE) to Carnegie Mellon University and the University of Pennsylvania, as well as NSF ITR Grant 0324883 to CMU and Stanford for classroom video databases. Current support comes from the NSF SCOTUS grant, the NSF PSLC grant, and NIH Grants to CMU for CHILDES, PhonBank, and AphasiaBank.[438]

While CHILDES has been developed and is based in the USA, the CHILDES database is openly accessible to anyone with a Web connection; contains over 44 million words (correct in 2007); has been referenced in over 2000 published articles (correct in 2003); and has over 4500 subscribers.[439]

All the data deposited into the CHILDES database are formatted in the style of the CHAT transcription system.[440] CHAT is described as

> a standardized format for producing computerized transcripts of face-to-face conversational interactions. […] It can be used with learners of all types, including children, second-language learners, and adults recovering from aphasic disorders.[441]

As a result of this data standardisation, all these CHAT formatted data can be analysed by the CLAN programs and other interoperable external data analysis programs.[442] Therefore, the FLLOC and SPLLOC researchers use the CHAT transcription system.

CHILDES exerts a strong influence over the provenance metadata, legal, technological and socio-cultural frameworks employed by FLLOC and SPLLOC, which are detailed in the subsequent three sections of this chapter.

---

[436] *Ibid*; Brian MacWhinney, 'The CHILDES Project: Tools for Analyzing Talk – Part 2: The CLAN programs', *CHILDES Manual*, Carnegie Mellon University (23 June 2015) <http://childes.psy.cmu.edu/manuals/CLAN.pdf> [accessed 9 August 2015].

[437] *TalkBank Website* <http://talkbank.org/> [accessed 9 August 2015].

[438] *Ibid.*

[439] MacWhinney, 'The CHILDES Project: Tools for Analyzing Talk – Part 1: The CHAT Transcription Format', p. 14.

[440] *Ibid*, p. 9.

[441] *Ibid*, p. 14.

[442] *Ibid.*

### 6.1.5 FLLOC and SPLLOC's legal framework

As a result of their membership of CHILDES, FLLOC and SPLLOC adhere to the TalkBank ground rules, which comprise eight key areas:

1) *Basic rules* for data usage.
2) *Disclaimers* regarding the use of data and programs.
3) *Principles* of data-sharing.
4) Talkbank policies for *data preservation* and *data workflow*.
5) *Options* for data-sharing.
6) *Guidelines for IRB consent forms*
7) The TalkBank *Code of Ethics* for sensitive data
8) How to submit *new data*[443] [Numbering and italicised emphasis from original webpage.]

Basic rules for data usage (CHILDES' ground rules) makes it clear that all data made available through CHILDES are governed by a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence; a brief overview of Creative Commons is provided in Chapter 2, section 2.3. Under the conditions prescribed by this licence, the human-readable summary outlines the permitted and excluded acts of the research users:

> Share — copy and redistribute the material in any medium or format [/] Adapt — remix, transform, and build upon the material [/] The licensor cannot revoke these freedoms as long as you follow the license terms. […] Attribution — You must give appropriate credit […] [/] NonCommercial — You may not use the material for commercial purposes. [/] ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original. [/] No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.[444]

While it is a strength that FLLOC's legal framework is built around a well-known licence provider, there is no direct mention of this Creative Commons licence or any other legal aspects on the FLLOC website (apart from attribution). There is a tab on the website specifically dedicated to ground rules, but this contains a hyper-link to the TalkBank website only. Whilst the first ground rule (basic rules for data usage) does reference Creative Commons, it does not state the type of licence. By providing a link to the webpage of the Attribution-NonCommercial-ShareAlike 3.0 Unported licence only, the legal framework remains buried under a number of Web pages.

---

[443] 'Ground rules', *TalkBank Website* <http://talkbank.org/share/> [accessed 9 August 2015].

[444] 'Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0)', *Creative Commons Website* <http://creativecommons.org/licenses/by-nc-sa/3.0/> [accessed 9 August 2015].

As the type of licence used is not brought to research users' immediate attention on the FLLOC website, this is potentially problematic as the specific permissions and constraints granted may be overlooked by research users. This could be easily rectified by providing a link to the licence utilised on the webpage of the datasets. On the FLLOC homepage, there are embedded icons to show that the FLLOC website is compliant with W3C standards; it could be useful if the Creative Commons icon were to be embedded also.[445]

### 6.1.6 FLLOC and SPLLOC's technological framework

The majority of the FLLOC and SPLLOC data are available from the websites in XML, which as an open, universal and non-proprietary format. Anyone is able to re-use that data with ease and without having to download bespoke and proprietary software.[446] Due to the widespread application and support of XML, this further safeguards the long-term preservation of these FLLOC and SPLLOC data.

Alongside the technological framework arising through CHILDES (CHAT, CLAN and the database) which has been discussed previously in the CHILDES overview, the FLLOC and SPLLOC websites offer another layer to this framework. Both of these websites and databases were developed by iSolutions which are the University of Southampton's professional IT services.[447] Alongside the interview participants' insights into the use of CHILDES, a grey area to be further addressed is the level of maintenance that the FLLOC and SPLLOC websites have received since their creation.

In contrast to the previous two case studies, the provenance metadata are presented in a more basic form of prose, and embedded within various paragraphs on a number of different Web pages. While these provenance metadata are extremely detailed, explaining the corpora and the collection of data, they are not as technologically sophisticated as the provenance metadata frameworks utilised by

---

[445] 'Mark-up validation service', *W3C Website* <http://validator.w3.org/check?uri=http%3a%2f%2fwww%2eflloc%2esoton%2eac%2euk%2f> [accessed 9 August 2015].

[446] 'FLLOC and XML', *French Learner Language Oral Corpora (FLLOC) Website* <http://www.flloc.soton.ac.uk/xml.html> [accessed 9 August 2015].

[447] 'Current team', *French Learner Language Oral Corpora (FLLOC) Website* <http://www.flloc.soton.ac.uk/team.html> [accessed 9 August 2015].

MEDIN, eCrystals and LabTrove. This area requires further probing by the interview materials.

### 6.1.7 FLLOC and SPLLOC's socio-cultural framework

While it is clear that FLLOC and SPLLOC were subject to rigorous ethics review, with their numerous projects ethically approved by the University of Southampton and other involved higher education institutions, there is not enough information given on the websites to reveal any of the potentially positive and negative experiences faced by the researchers in obtaining consent, upholding confidentiality and checking the quality of data. Moreover, there is no information on how the inclusion of more vulnerable participants (the school children and the young adults who were largely living abroad for the first time) and participants discussing very sensitive issues impacted on such ethical procedures. Were there any extra measures that needed to be taken? This is another grey area that requires further explanation from the interview participants, and is the core focus for this chapter.

## 6.2 FLLOC and SPLLOC: people selected for interview

**Dr O.** is a second language acquisition academic within modern languages at the University of Southampton. He has been involved with FLLOC and SPLLOC [$O_1$]. **Dr O.** was selected principally for his data management expertise.

**Dr P.** is a second language acquisition academic within modern languages at the University of Southampton. She has been involved with FLLOC. [$P_1$] **Dr P.** was selected principally for her data management expertise.

**Dr K.** is an interdisciplinary academic within modern languages at the University of Southampton, researching across the domains of life histories, linguistics and history. While she does not have direct experience with the FLLOC and SPLLOC projects, she was chosen to widen the scope of research in modern language domains as personalised and (in some cases) sensitive data are used in areas other than language acquisition research [$K_1$]. **Dr K.** was selected principally for her academic research expertise.

**Miss L.** is an academic liaison librarian at the University of Southampton. As part of this role, she runs workshops and events that address data management practicalities and issues across the sciences, the social sciences and the humanities. Although she does not have a direct connection with the FLLOC and SPLLOC projects, she was

selected for her extensive data management expertise, as none of these projects had a specified data manager. [$L_1$] **Miss L.** was selected principally for her data policy expertise.

**Mr M.** is a legal advisor within Research and Innovation Services at the University of Southampton with wide-ranging legal expertise, including intellectual property law, data licensing, ethics, and the digital open access and open data movements. Research and Innovation Services is the support service for University staff and students, offering guidance on how to release scholarly materials responsibly. They were involved in drafting the University of Southampton's data management policy. [$M_1$, $M_2$] **Mr M.** was selected principally for his legal expertise.

**Mr D.** is a member of the Technical Innovation Team within iSolutions, which is the University of Southampton's professional IT services. While he does not have direct experience with the FLLOC and SPLLOC projects, he has extensive experience of data archiving and re-use, and works with academics across the University to provide research support systems. [$D_1$, $D_{10}$] **Mr D.** was selected principally for his technological expertise.

**Mr D., Dr K., Miss L., Mr M., Dr O.** and **Dr P.** were interviewed during 2013. The interviews were approved by the University of Southampton's Management Ethics Committee and adhered to the planned methodology (see Chapter 3 for ethics notice and other information).

## 6.3 FLLOC and SPLLOC: interview materials

### 6.3.1 FLLOC and SPLLOC's provenance metadata issues

#### 6.3.1.1 Read-me files

As with the previous two case studies, provenance forms an integral part of the FLLOC and SPLLOC projects and yet again the interview participants describe provenance in a very positive manner, as important [$D_{12}$, $M_9$, $O_{14}$], crucial [$L_{11}$] and critical [$P_{18}$].[448]

[448] **Dr O.** states:

> […] Obviously [provenance] it's really important, because if […] you've got transcripts up on a website, and they're there by themselves – it doesn't mean anything to anybody. You don't know […] how they were collected, who collected them, when they were collected. […] [$O_{14}$]

Provenance is not only important for establishing data origins [$D_{12}$], but for rights clearance too [$M_9$].

The specific parameters of provenance metadata depend on its contextual basis, and therefore it is the responsibility of the researcher who collects a dataset to provide all necessary provenance information [$L_{11}$]. Necessary provenance information includes the metadata that makes a dataset understandable; prevents research users constructing a false assumption about a dataset; and, facilitates trust and valuable re-usage [$L_{11}$]. The researchers involved with the FLLOC and SPLLOC research projects ensured they recorded all crucial provenance metadata.

At the time of the interviews, the LANGSNAP Corpus data were not yet openly available, nor were their provenance metadata.[449] **Dr P.** therefore explains the key provenance metadata attached to their transcripts [$P_{18}$]. The provenance metadata includes: the data collection point; date of the interview; the identification number issued by LANGSNAP and gender of the interviewers and participants; the country the data were collected; and, the particular phase (there are six in total) in which the data were collected [$P_{18}$].[450]

It is clear that the read-me file [$O_{14}$] style of provenance metadata utilised by the FLLOC and SPLLOC projects is not as technologically sophisticated as the provenance metadata frameworks employed by MEDIN, eCrystals and LabTrove. FLLOC and SPLLOC impart a new perspective on provenance metadata not found in the case studies of the previous two chapters. The provenance metadata prove to be more verbose in comparison to other standardised metadata formats (such as, Dublin Core and GEMINI). Without summarisation and separation of the provenance metadata into key fields, these data may be less discoverable. However, the success of academic research data re-usage does not necessarily lie in the sophistication of the technology that underpins the provenance metadata. Although it is often assumed that the more complex the IT aspects of academic research the more searchable and re-usable the academic research data becomes, it is the quality of the metadata that is of primary concern. Although enhanced machine-readable (or machine-understandable)

---

[449] **Dr O.** confirmed the type of provenance metadata available on FLLOC and SPLLOC, as described in the evaluation of the primary source materials. FLLOC and SPLLOC have the same standardised provenance information [$O_{15}$] in the form of word documents (read-me files) that accompany the transcripts and sound files on the corpora websites [$O_{14}$]. These word documents outline: the research; the research questions; who the participants were; what methods were used to collect the data – such as spontaneous conversion or a particular task; and, 'any issues that there might be with the data *um* in terms of *um* re-use' [$O_{14}$]. There is a whole metadata set on the FLLOC and SPLLOC websites which further describe: about the project; the project rationale; who the learners were; who the research team are; and the tasks used [$O_{14}$].

[450] LANGSNAP employed a team of transcribers and anonymised their details [$P_4$].

provenance metadata would aid discovery (for example via search engines) and more succinct categorisation (such as the metadata fields used in the case studies of the earlier chapters) would offer a more concise way to attain the key metadata, this is a secondary concern. Therefore, FLLOC and SPLLOC demonstrate that the creation of robust metadata comes first; their organisation and machine readability, although important, are secondary.

Standardised metadata formats are not without their problems however [$D_{14}$]. For instance, Dublin Core standards can lead to the creation of very restricted metadata, as there are only a very small set of fields [$D_{14}$].[451] Dublin Core provides bibliographic metadata, but does not satisfy the more extensive definitions of provenance (e.g. the thesis definition of provenance, traditional definitions of provenance and PROV-O definition – see Chapter 2, section 2.3.4). As a result 'there'll be huge numbers of items that share very similar Dublin Core metadata sets' [$D_{14}$]. While metadata standards, such as Dublin Core, set useful minima that can be modified by future users, overly standardised metadata can cause a significant issue for data discovery. Moreover, there is yet to be consensus on the best representation for publications (for instance the variety of referencing systems available from Chicago and ACM to Harvard and MHRA). Therefore, it is unsurprising that there is also yet to be a unified approach to data and metadata [$D_{14}$].[452]

As personalised and sensitive data must be managed ethically, provenance metadata pertaining to such data also require the same level of maintenance. The LANGSNAP research team ensured that none of the participants' personal or sensitive data would be exposed through the provenance metadata by providing the location of the interviews at the level of a country [$P_{18}$]. For illustrative purposes, if an interview was conducted in the city of Salisbury, it would appear in the provenance metadata as the UK only [$P_{18}$].

**Miss L.** raised a further issue not covered in the two previous chapters. Provenance metadata are not only for the benefit of other/future research users, but are

---

[451] **Mr D.** further raises a point about how to describe the individual who collects a dataset – should they be described in the provenance metadata as: a collector, a collator, a generator, an originator or a data owner [$D_{14}$]?

[452] **Mr D.** does not believe, therefore, that a universal provenance metadata standard can be created for all academic research data [$D_{14}$]. **Mr D.** contends:

> […] The way to handle that is not to insist on a standard. […] Standards will grow organically – they always do. People that care about certain sets of datasets will come up with a standard that works for that set of datasets. […] [$D_{14}$]

an essential record for the researchers who originally collected the data [$L_{11}$]. No researcher is able to recall every exact detail about a dataset in the coming years without provenance metadata, as **Miss L.** explains:

> […] if people say: 'oh why do I have to do all this – *erm* metadata stuff. What – what's that all about?' […] They're actually looking at it – *erm* from the point of view of someone who's so intimate with that particular set of data that they never think they will ever forget how and why it was created. […] [$L_{11}$]

Through relying on their provenance metadata to elicit the exact times of when data were collected, the LANGSNAP research team were able to assess the participants' language development over a determined duration [$P_{18}$]. Provenance metadata, therefore, is able to safeguard the usage of data by researchers during the duration of a research project, and long after project completion [$L_{11}$].

### 6.3.2 FLLOC and SPLLOC's legal issues

#### 6.3.2.1 Protection of personal and sensitive data

The researchers involved with FLLOC and SPLLOC signed an agreement with CHILDES to ensure that any uploaded data would be openly accessible to anyone with a connection to the Web, and to deposit a copy of these data within the CHILDES database [$O_{10}$]. This decision was in part due to the nature of the FLLOC and SPLLOC data collected. While participants were children and young adult learners there were no further levels of sensitivity to consider, which would require limiting access, as is the case with other data held within the CHILDES database. Alongside data on child language, bilingual language and second language acquisition data, there are language disorder data and videos of language on the CHILDES database that understandably require a much stricter permissions process. This is, therefore, a strength of CHILDES, as the developers have taken into account the varying degree of sensitivity/protection pertaining to certain types of data, and recognition that not all data can be released as open data.

The Data Protection Act 1998 is also recognised as a vital statute for personal data, and is emphasised throughout the ethics process [$L_8$]. However, to what extent researchers think about this in strict legal terms or within ethical conventions is a grey area. This further raises the issue about whether databases based in other jurisdictions (such as CHILDES based in the USA), and their international research users, actively comply with minimum legal standards enforced by the UK and EU in instances where

academic research data are gathered during the course of UK research. Which legal jurisdiction takes precedence over the re-usage of these academic research data?

#### 6.3.2.2 Awareness

The FLLOC and SPLLOC websites do not reference the Creative Commons licence selected by CHILDES under which the data held are made available. The interviews with two of the researchers directly involved with the Young Learners Corpus and the LANGSNAP Corpus did not help to clarify this issue. This was because, by the researchers' own admission, they had limited awareness of legal issues [$O_9$, $O_{10}$, $P_{17}$] and they had not personally worked with Legal Services or Research and Innovation Services in conjunction with these projects [$O_{11}$, $P_{15}$]. However, **Dr P.** was certain that a member of FLLOC and SPLLOC would have sought legal advice at some stage [$P_{15}$].

While the interviewees were unable to expand on the specific licensing issues, this lack of awareness exposes another key grey area for legal best practice – is it the researcher's responsibility to understand all the legal issues? As was shown in the eCrystals and LabTrove Case Study and re-iterated by **Mr M.**, some researchers are 'keen that nothing gets in the way of their research' [$M_7$]. Although this does not appear to be the case with FLLOC and SPLLOC, as CHILDES has selected the Creative Commons licence and another member of the team would have sought legal advice, it is interesting to note whether researchers are responsible for permissions. **Dr K.** considered that ethical issues must be the focus of the researcher and the higher education institution and/or archive is better placed to advise researchers of the legal issues.

### 6.3.3 FLLOC and SPLLOC: technological issues

#### 6.3.3.1 Formats

The primary source materials revealed that, from the 1950s onwards, the discipline of second language acquisition research (and other research areas that collect oral data, such as life histories) has been strengthened through the development of digital recording technologies. Even in the short time since the Linguistic Development Project in 2001, recording equipment has largely improved the upload of data, the quality of video and sound [$O_{22}$]. A number of technological changes (including recording and the

Web) have occurred over the life-span of FLLOC, and therefore **Dr O.** contends that technology is the biggest influence on second language acquisition research [$O_{22}$].

For instance, the first FLLOC project – the Progression Corpus – collected data from secondary school pupils between 1993-6.[453] These data were originally recorded on cassette tapes that were later digitised and uploaded to the FLLOC website [$P_{33}$]. Therefore, the FLLOC researchers have taken measures to reformat their data and preserve it for future re-usage.[454] Furthermore, the CHILDES developers have digitised considerable amounts of data collected from child language research projects conducted in the 1960s and 1980s [$O_4$]. FLLOC is an example of a sustainable academic research data re-usage model that has adapted to a number of technological changes over its lifespan.[455]

The Young Learners Corpus and the LANGSNAP Corpus do not use bespoke software and provide their data in non-proprietary formats, such as XML [$P_4$].[456] Therefore, research users are not required to download the freely available CHILDES software to access the recordings and transcripts [$P_4$]. Moreover, this is beneficial from a cost point of view, as research users do not have to spend extra money on proprietary software [$P_4$].

While CHILDES offers a number of community supported and global standards for transcription, the LANGSNAP research team still faced difficulty with utilising the same conventions across French and Spanish [$P_4$]. **Dr P.** offered the following example: whereas the Spanish subjunctive is clear and obvious, in contrast it is difficult for researchers to distinguish between the present indicative and present subjunctive with some French verbs [$P_4$]. An autographic transcription is uploaded to CHAT as

---

[453] 'Progression Corpus: Description', *French Learner Language Oral Corpora (FLLOC) Website* <http://www.flloc.soton.ac.uk/progression/index.html> [accessed 9 August 2015].

[454] Format shifting is a very important aspect of modelling best practice for academic research data across the sciences, the social sciences and the humanities. **Miss L.** offers an interesting example:

> […] I heard about yesterday about someone who took observations on a BBC Micro in 1980. And they have valued that information sufficiently to have saved it as ASCII text, and to convert it from BBC Micro Floppy Disc … through the various different standard formats – so that there is now a PhD student – updating, working with that data. How many years is that? – Thirty plus years later. […] [$L_4$]

[455] **Miss L.** believes researchers should be wary over the 'next greatest thing to come' – as shown by Betamax vs. VHS [$L_4$]. There is a balance between new technologies becoming the standard and others that fail to become popular [$L_4$]. Researchers should decide on data formats at the beginning of their research project – 'format is such an important *um* choice, and also managing it' [$L_4$].

[456] **Miss L.** maintains that individuals who write bespoke programs without publishing the accompanying software or script may cause potential problems for future preservation [$L_4$]. Furthermore, it is a necessity that some researchers who utilise propriety brand software use the save as function to output results in non-propriety formats such as .txt, .csv or .odt [$L_4$].

representation of the interview, but does not make distinctions between subjunctives, therefore LANGSNAP further employed a speech tagger to make this clear [$P_4$]. It is evident that although standardisation is beneficial there is no one-size-fits-all policy (in line with the findings of the previous two chapters) and such standards have to be tailored to suit certain contexts.

#### 6.3.3.2 Discovery

The FLLOC and SPLLOC corpora both receive many visits from research users – FLLOC has around one hundred hits per month [$P_{25}$]. FLLOC and SPLLOC are described by **Dr O.** as quite unique, as they focus on a very specific context – schoolchildren learning languages within the British school system [$O_4$].[457] Rather than traffic flowing directly from CHILDES to FLLOC or via the (serendipitous) use of a search engine, it appears that data discovery is sustained through a textbook written by Professor Rosamond Mitchell and Professor Florence Myles about language learning [$P_{25}$]. Therefore, future advances such as texts and articles can add value to the discoverability of academic research data.

The discoverability of FLLOC and SPLLOC data is quite different to the data held by MEDIN, which is reliant on an authoritative, digital discovery metadata catalogue. There appears to be no such robust and centralised discovery portal for FLLOC and SPLLOC. While CHILDES maintains a sizable record of global child language data projects, it does not seem to be facilitating data discovery on the same scale as MEDIN. For the FLLOC and SPLLOC data, discovery is facilitated largely through the print context of this research (the published textbook) and the reputation of the researchers involved. However, it is unclear whether this discovery is enabled primarily by: (a) the importance of this project within a particularly unique contextual framework; (b) the reputation of the people behind the project regardless of the data in a number of research communities; or (c) just because it is openly accessible. This raises a further grey area – to what extent would data discovery be increased through the creation of a data discovery portal? How can the value of a project and its data be extended beyond a reputational lifespan?

[457] **Dr P.** maintains that no two datasets collected from participants in second language acquisition research will be the same due to their inherent human element; additionally, some participants provide lots of data whereas other individuals only offer brief information [$P_{26}$].

As was shown by the eCrystals and LabTrove case study, integrity of identifiers on the Web is a vital aspect of data discovery and sustainable re-usage [$D_6$]. However, FLLOC and SPLLOC do not utilise persistent identifiers, such as DOIs, and these websites and their URIs fall under the University of Southampton's domain.[458] While it is not necessary for researchers to rely on persistent identifiers [$D_6$], it appears that the current and future reliability of the FLLOC and SPLLOC URIs is dependent on the University of Southampton's management of URI integrity [$D_6$]. It is unclear to what extent these current identifiers will be maintained over time – and this is a significant grey area for individuals and organisations modelling best practice for academic research data re-usage for the longer term. Broken URIs that impede data discovery must be prevented and this issue is being reviewed by the University of Southampton (for example the DataPool project, see Chapter 5: eCrystals and LabTrove Case Study) [$L_7$].

#### 6.3.3.3 Technological support

The interviews confirm that CHILDES is extremely popular worldwide [$O_4$]. The CHILDES developers do not have their own research agenda [$P_{29}$] but constantly strive to improve the software for the research community through releasing new software versions and updates [$P_{29}$]. The open, active email list is an extremely important aspect of technological support for researchers involved with FLLOC, SPLLOC and other child language data projects [$O_4$, $P_{29}$].

By simply receiving emails from the subscribers directly, the CHILDES research developers are able to respond rapidly to any issues faced by subscribers [$O_4$, $P_{29}$]. For instance, during work on the LANGSNAP project, **Dr P.** encountered such a technological issue where she was unable to transfer her interview sound recordings into NVIVO, because they were in an incompatible format [$P_{29}$]. After emailing the CHILDES developers, they rectified the problem swiftly by creating a command that converted the interviews into a compatible format [$P_{29}$].

Due to the Web and its role as a global distribution network, research users have access to greater technological support than ever before. Researchers from the

[458] Although not directly relevant to this case, academic research data with a specific university's URI is indicative of quality, such as any URI under the ac.uk domain [$D_7$]. **Mr D.** maintains that relying solely on such a URI potentially is a false authority however [$D_7$]. **Mr D.** states: 'no university is really completely aware of every single document that's under its domain' [$D_7$].

University of Southampton and beyond are able to contact research developers based at Carnegie Mellon University in the US with their problems immediately (subject to time differences). This level of open technological assistance is a strength of using a software provider such as CHILDES.

The researchers involved with FLLOC and SPLLOC are capitalising on their knowledge of CHILDES, and training other researchers about this software through workshops [$P_{28}$]. Therefore, secondary technological support from other researchers is useful to further enhance uptake and promote awareness of robust technological systems (such as is the case with the MEDIN core team in Chapter 4, who actively offer support to data depositors).

The responsibility for the functionality and organisation of the websites and databases lies with the developers rather than the project researchers [$P_6$]. From FLLOC and SPLLOC's primary source materials, it is clear that the researchers did not rely on technological assistance from the CHILDES developers only, but required help from iSolutions at the University of Southampton to create the FLLOC and SPLLOC websites and databases. However, such as is the case with many higher education institutions, the University of Southampton does not automatically provide IT support to research projects [$P_{19}$]. It is therefore important that researchers, particularly with limited technological knowledge, plan for IT support, as good academic research data re-usage requires secure technological underpinning. Funding was set aside to employ an IT specialist from iSolutions to develop the websites and databases for the Young Learners project and the LANGSNAP project [$P_6$]. Further collaborative work with IT professionals in the area of life histories to develop digital methodologies would be beneficial also [$K_{22}$].

While FLLOC, SPLLOC and the two projects have received robust assistance from iSolutions over their course, as with many studies there is no funding set aside to cover future maintenance of the websites and databases after project completion [$O_6$]. Therefore, this leads to a further grey area for modelling best practice. Without continual maintenance from IT professionals these data and many others may be jeopardised. How sustainable is the long-term storage, updates to and archival of academic research data?

While there are copies of these data within the CHILDES database that offers a secondary way of long-term preservation, the FLLOC and SPLLOC websites appear to be the most popular platform for re-usage and therefore need to be sustainable.

However, who is able to and should take responsibility for this ongoing maintenance – the higher education institution? In the growing myriad of data and the limited resources available is this currently an option? Moreover, as researchers retire, change jobs and leave employment who should have overall responsibility– the institution or the researcher? However, as some higher education institutions do not permit researchers to take copies of academic research with them elsewhere – is this an option?

#### 6.3.3.4 Data analysis

Due to the growing number of second language acquisition research corpora, researchers are able to (re)use data from hundreds of language learners leading to more representative findings [$O_{22}$]. This is a dramatic change from the recent past where, typically, researchers were able to (re)use data from a couple of language learners only [$O_{22}$]. Due to this greater data availability, more powerful software has emerged to help researchers analyse these data by deriving further data from sound recordings, such as the frequency of specific terms [$O_{22}$].

CHILDES adds value to the data it holds by enabling researchers to access and interrogate corpus data, for example through frequency lists (the number of times a specific term is used within a sound recording); and collocations which is defined by the OED as 'the habitual juxtaposition of a particular word with another word or words with a frequency greater than chance' [$P_{31}$].[459]

The LANGSNAP research team uses additional software for quantitative analysis and to derive further data [$P_4$]. The researchers used ELAN for quantitative analysis – and to manually count the syllables within the sound recordings [$P_5$]. This software was developed and maintained by a team of technologists, developers and linguists at the Max Planck Institute for Psycholinguistics.[460] Similar to CHILDES, ELAN is open to the global research community and supported by a team of developers.

Originally, to help with speech analysis, the LANGSNAP researchers considered using the Doing Phonetics by Computer (Praat) software developed by Paul

[459] 'Definition of collocation', *Oxford Dictionaries Website* <http://www.oxforddictionaries.com/definition/english/collocation> [accessed 9 August 2015].

[460] 'The Language Archive: Team – Expertise', *Max Planck Institute for Psycholinguistics Website* <http://tla.mpi.nl/team/expertise/> [accessed 9 August 2015].

Boersma and David Weenink at the University of Amsterdam to derive this data [$P_4$].[461] However, the Praat programme seemed unsuitable in the circumstances, as the data were being collected in public places such as restaurants [$P_5$]. As Praat calculates the number of syllables in a sound recording by focusing on speaking time, silent time and pauses, the background noise makes it hard to distinguish silences [$P_5$]. Therefore, reliability of the data in these circumstances is not enough; the quality of the sound recordings can impact on further re-usage [$O_3$].

Given all the various open (and supported) software solutions available, it still remains difficult to change some researchers' familiar and established practices in order to enrich the data analysis process. For instance, there are still a number of second language acquisition researchers who will only use Microsoft Word to derive quantitative data from transcripts [$P_{31}$]. This is mainly through the ctrl+f search function, which is extremely limited [$P_{31}$] e.g. it is very time-consuming to extract all the collocations within a transcript with ctrl+f alone. Therefore it is not enough that these various open (and supported) software solutions are available, (as was shown in the MEDIN case study) to encourage and increase their utilisation researchers need to be shown why they are beneficial.

The LANGSNAP researchers used iSurvey, an online and open questionnaire platform developed at the University of Southampton to capture qualitative data about learners' social networks and personalities [$P_4$]. From this, it is clear that the LANGSNAP researchers expanded their technological framework with open and supported software to add further value to their data, and this is, therefore, a good example of best technological practice.

### 6.3.4 FLLOC and SPLLOC: socio-cultural issues

#### 6.3.4.1 Difficulties with disclosure

Academic research data gathered from human participants are among the most difficult to release or disclose fully as open data.[462] Researchers need to be cautious where the

---

[461] **Dr P.** [$D_1$]; *Doing Phonetics by Computer (Praat) Website* <http://www.fon.hum.uva.nl/praat/> [accessed 9 August 2015]; Paul Boersma and David Weenink, 'Praat: doing phonetics by computer [Computer program]'. Version 5.4, *Doing Phonetics by Computer (Praat) Website* <http://www.fon.hum.uva.nl/praat/> [accessed 9 August 2015].

[462] **Mr D.** is very supportive of the open data movement:

data are personal and/or sensitive [$O_{23}$], and therefore aware of circumstances where the distribution of such data is likely to cause harm [$D_{21}$]. Such as in certain instances where data: are obtained from individuals who are at risk from death, dismissal and ridicule – for example dissidents and whistle-blowers [$D_{21}$]; from vulnerable individuals – for example children [$O_{23}$]; involve discussion of physical and emotional abuse [$L_{14}$]; and, may contain views that would otherwise be censored as, for example, racist, sexist, violent and homophobic [$L_{14}$].[463] Sensitive data cannot be treated the same as recording environmental observations [$L_{14}$].

Higher education institutions support the release of open data where it is necessary and practicable [$M_3$]. However, some participants will not permit sound recordings and transcripts to be made openly available even where pseudonyms and other confidentiality measures are employed [$K_6$]. An individual's control and consent over re-usage of their personalised data are important and must be respected [$D_{21}$]. The research methodology needs to explicitly state why and how sensitive data are vital to the project [$L_{14}$]. Moreover, it needs to outline in what ways data are to be handled; data managers require assurances that such data has been ethics approved [$L_{14}$]. Therefore, it is vital that researchers adhere to ethical guidelines and procedures when gathering and (re)using data from human participants [$P_{12}$]. The ethical guidelines and procedures prescribed by researchers' higher education institutions and CHILDES are essential and

---

> […] I think what we're doing with open data is we're changing the world. So, so … there's never been a time in history when so many people had so much access to so much information. […] If you want to do research over datasets – some datasets are available already – it's only going to be – get better when, when, when more datasets become publically available. And this is […] a good with a capital G. […] [$D_{20}$]

[463] Alongside ethics issues, there are further reasons that prevent the release of open data including: (i) trade secrets – nobody knows about the data outside those involved with the research project such as where a company pays for that privilege; (ii) its release would constitute a (potential) detriment to national or international security; and, (iii) a controversial research area where some individuals believe a research project is abhorrent or unacceptable [$L_{12}$]. Moreover, (although not directly relevant to this chapter) there are other examples of academic research that are sensitive, but do not involve human participants. For example, researchers need to consider the broader impact of their research – especially when their data comes from hazardous materials [$M_{10}$]. It is difficult to navigate how such academic research data will be re-used – such as the tobacco industry re-using data for commercial purposes, and political regimes re-using data for weapons of mass destruction [$M_{10}$]. **Mr M.** asserts that researchers must think beyond a UK audience, and remember that the Web has a global audience [$M_{10}$]. This issue relates to a minority of academic research data only – therefore, researchers should not become paranoid over the re-usage of their data [$M_{10}$]. In relation to this, **Mr M.** raises an important legal issue – the Export Regulations that include the transmission of data out of country [$M_{10}$]:

> […] In many instances [Export Regulations] this will not be an issue. However, there is data related to potentially nasty biological *um*, *um* samples – and potentially, you know, high powered laser – anything that could get into weapons of mass destruction *um* and/or facilitate their use – that I'm not entirely sure *um* really goes through most researchers' minds. *Um* and I know that very few universities really have tackled or understand this piece. […] [$M_{10}$]

integral components of FLLOC (and SPLLOC) [$P_{12}$]. However, as not all UK higher education institutions have the same ethics approval (procedures are not standardised) – how can a level of national consistency be recognised? Moreover, where data collection occurs as a result of a joint project with an external partner, which ethics procedure takes precedence? Who is responsible for checking ethics and any incompatibilities – the research ethics office or the principal investigator?

#### 6.3.4.2 Consent

Consent forms now constitute an essential ethical basis for research conducted with human participants [$K_4$, $O_{22}$], as they record the mutual understanding between researcher and participant [$K_5$]. By establishing a written agreement, a consent form imposes a strict set of parameters that outline how the collected data will be used, published and maintained by the researcher and (potentially) re-used by other research users [$K_{10}$, $P_{12}$].

As the Web extends the distribution network of data to reach a (potential) instantaneous global audience, consent has become a bigger issue than that experienced in the age of print. For instance, in the 1970s it was normal for researchers not to use consent forms when conducting research with human participants [$K_4$]. **Dr K.** admitted that despite her role as an interviewer in the field of life histories research, she was uneasy about being interviewed until she read all the interview documentation, and was able to acknowledge the anonymisation procedures used [$K_{24}$]. This is because she realises: 'how easy it is for digital data to be transmitted *um* and used in all sorts of different ways nowadays' [$K_{24}$]. Consent forms, therefore, are a vital part of the socio-cultural framework, and are viewed largely in a positive manner by the interview participants [$O_{22}$].

This instantaneous, global network has sparked a confidentiality and/or privacy paradox. On one hand, participants may be more reserved (therefore producing more controlled and/or limited data), or individuals may decline an invitation to contribute to research altogether. For instance, while open access has long been at the centre of life histories research by providing a channel for often subaltern and/or minority groups whose narratives would otherwise remain hidden, the Web potentially stifles this research as participants are increasingly reticent about participation [$K_{20}$] or concerned about sound recordings and transcripts being made openly accessible [$K_6$].

On the other hand, younger generations whose lifetime has been dominated by the Web – in particular social media – appear to have lower expectations of privacy [$L_{14}$].[464] Due to the nature of social media, it is common for individuals, who are perhaps less concerned about their privacy, to disclose personal information, such as dates of birth, political preferences, relationships and exchanges with other users [$L_{14}$]. In the future it is possible that current socio-cultural frameworks may be re-shaped further by researchers and participants' shifting expectations of privacy [$D_{23}$, $L_{14}$]. Will this change responsibilities and liabilities if there is an ethical breach?

The reliance on consent forms and ethics procedures make data collection more difficult in a digital age [$O_{22}$]. An ethical issue arises where a participant refuses to sign a consent form, but verbally agrees to take part in the research and for such data to be openly disseminated [$K_3$]. **Dr K.** offers an example of a participant who, as an anarchist, refused to sign a consent form for ideological reasons [$K_3$]. While the data from this interview could be included in the original research, it could never be deposited for public re-usage in a paper or digital archive due to the lack of a consent form [$K_3$]. Further to this, issues arise where participants may withdraw altogether when they find out that they have to sign a consent form, or they realise that they will not be paid for their data.

Although the substantial number of policies, guidance and completed forms that are required for ethics clearance provide a robust provenance trail, all these prescribed criteria have the potential to stifle the collection and re-usage of academic research data [$K_{19}$]. Rather than directly considering the academic beneficial or negative implications of deviating from the predetermined ethics procedure, researchers may take a format driven approach and aim to tick all the boxes [$L_8$]. This can result in researchers going beyond mandatory requirements set by their higher education institution, funding bodies and legal frameworks, leading to a situation where data are not re-usable by researchers or the actual individual collecting the data [$L_8$]. Again, as illustrated by the MEDIN case study, there needs to be a degree of flexibility [$K_{19}$].

---

[464] **Mr D.** states:

> […] I think all of the big explosions have happened on the internet. I think we're coming to terms with it, and how to use it. […] As younger generations come into academia – the gen – the Facebook generation comes in – they're not going to ask: 'why should I share my data'? They're going to ask: 'why can't I share my data?' And I think that's a sociological rather than technological change. But I think it's the sociological changes that are more powerful. They leverage technology. […] It's a question of generations – that the current academic generation is perhaps not comfortable with the current state of the art of technology. […] [$D_{23}$]

#### 6.3.4.3 Ethics procedures

As the Young Learners Project involved researchers from both the University of Southampton and the University of Newcastle, this project was assessed and approved independently by the ethic boards at these two respective higher education institutions [$O_{17}$]. Robust ethics procedures were important [$O_{17}$], as not only did this research involve human participants – but minors. Therefore, the whole research team was required to have Criminal Record Bureau (CRB) checks [$O_{28}$].

The researchers involved with the Young Learners project used consent forms to obtain permission from the schoolchildren's parents and/or guardians, and schools [$O_{17}$]. Where the participants were over sixteen (in sixth form), researchers also had these students sign a consent form [$O_{17}$]. School children were withdrawn from all classes that involved data collection where their parents and/or guardians (and participants over sixteen) had not given their consent for participation [$O_{17}$].

The LANGSNAP Project involved researchers from the University of Southampton and therefore this research approved by Faculty of Humanities' ethics board only [$P_8$]. However, the researchers followed a number of further ethical guidelines including: the British Association for Applied Linguistics (BAAL), and the ESRC [$P_7$].[465] Moreover, the researchers undertook mandatory training from the University of Southampton's counselling services [$P_{15}$].[466] This was because (potentially) the participants would be providing confidential information, for example, affecting their studies and could feel homesick [$P_{15}$].

In the same way the Young Learners project was peer-reviewed by the ESRC, the LANGSNAP project was scrutinised to ensure that the scope of the project, its methodology and outputs were feasible and thoroughly planned [$P_7$]. As with the Young Learners project, the researchers will therefore submit a final report and a database that will be peer-reviewed by the ESRC to confirm that the researchers have fulfilled the funding bid requirements [$P_7$].[467]

---

[465] *British Association for Applied Linguistics (BAAL) Website* <http://www.baal.org.uk/> [accessed 9 August 2015].

[466] 'Health and support', *University of Southampton Website* <http://www.southampton.ac.uk/undergraduate/studentlife/healthandsupport.html> [accessed 9 August 2015].

[467] Correct at time of interviews in 2013.

Ethical procedures are a necessity not only at the beginning of a project; they must be continually validated through to the completion of a project. The nature of research is that it develops over a study. In order for a project to retain ethical compliance all subsequent revisions need to be captured and undergo independent verification. In some instances, the University of Southampton's ERGO system requires that a form is submitted at the end of a research project that notifies the ethics board of any changes made in the course of research.

It is evident that the welfare of the individuals – interviewees and interviewers – is a vital consideration for research that involves human participants and an essential part of modelling best practice. The University of Southampton's RGO procedures explicitly require researchers to complete a risk assessment that includes an appraisal of their personal safety, for instance lone interviewing in an unfamiliar, private residence could be considered a potential high risk.

Higher education institutions and research councils have a vital role in the independent verification and robust checking of research projects and their methodologies through peer-review, (faculty) ethics boards, procedures and guidelines.

#### 6.3.4.4 Confidentiality

The data available through FLLOC and SPLLOC are anonymised. The researchers from the Young Learners project sought to remove all references to names of schools, children, teachers and family members that appeared in the sound recordings and transcripts [$O_{17}$]. For example, in the sound recordings all names were blanked out [$O_{20}$]. FLLOC and SPLLOC retain a record of all the participants' pseudonyms on a separate server, which only the researchers are permitted to access [$O_{19}$].

The LANGSNAP research team not only anonymise the participants' names, but any content that could possibly lead to their identification (such as factual information) or would be considered as potentially sensitive (such as stories about partners or family members) [$P_{12}$, $P_{26}$]. The utilisation of pseudonyms is a common feature of life histories research too [$K_6$].

While pseudonyms are an important tool to facilitate the preservation of confidentiality, some participants may wish their real names to be utilised within research [$K_6$]. This was the experience of **Dr K.**, where all but one of her participants wanted their real name used as their narratives had not been documented at the time [$K_6$]. Therefore, there will be a minority of research projects that do not use

pseudonyms, but this choice should be made by the participant and/or scoped as relevant in the rationale and ethics compliance for the project.

#### 6.3.4.5 Checking procedures

The responsibility for quality assurance appears to lie primarily with the researchers who collect and use the academic research data [$D_4$, $M_3$]. The Young Learners Project's research assistants were responsible for checking the sound quality of the audio data collected, and the associated transcriptions as reliable, robust, fit for purpose and compliant with CHAT (CHILDES' transcription procedures) [$O_6$]. Moreover, in the same way as the previous two case studies, independent checks were carried out by the research team [$O_6$]. For instance, the research assistant who transcribed a particular sound recording would not assure the quality of the resulting transcription, another research assistant would be responsible for carrying out those checks [$O_6$].

Although these data (from the Young Learners Project) were vigorously scrutinised, a minor number of post-project data corrections have been made by the research team to rectify errors or mistakes previously missed [$O_6$]. FLLOC and SPLLOC have not received any feedback from research users with regards to data quality however [$O_6$]. Again, this echoes the point made in the previous two chapters that academic research data are not error free in the majority of instances, and therefore research users need to be aware that there is always a risk of human error.

Where post-project data corrections were carried out, these appeared to be on a voluntary basis. This is a common issue across disciplines within the sciences, the social sciences and the humanities. Often the funding is not available to actively maintain the data and database after project completion [$O_6$]. Moreover, it is usual for members of the original research team to focus on subsequent research projects and eventually leave that higher education institution (perhaps for another academic position or due to retirement) [$O_6$].

As a result of the University of Southampton's data policy, significant data created at the University need to be preserved for at least ten years. The key issue for projects with no necessary continuity is whether this ten year time-frame is long enough. eCrystals currently differs from this as it is an ongoing project with no proposed and finite date of completion, and therefore long term preservation of data is perhaps not such an urgent concern. Another important issue is what constitutes

significant data under this University policy and therefore what should be safeguarded for future re-usage [$L_5$].

As was shown by the previous two case studies, published research and their underlying data have an intrinsic and important relationship [$D_9$].[468] While it is accepted that the peer-review process acts as the essential arbiter of academic validation and scrutiny [$M_6$] within second language acquisition research it appears that they do not check the underlying research data [$O_8$]:

> […] I do think that there is a very healthy peer-to-peer scrutiny that is done through the journal mechanism. *Um*, and there is certainly a very healthy *um* academic rivalry within specialists in certain fields. *Um*, and they should and will, *um* test *um* each other and the quality of their research. I believe that that is probably one of the best mechanisms of ensuring *um* … research integrity and quality. […] [$M_6$]

For instance, **Dr O**. has never been asked to provide any data (such as sound recordings and transcripts) to an academic journal publisher, because the default position of the publisher is to trust and have confidence in the data analysis and interpretation undertaken by the researcher [$O_8$].[469] **Dr K.** re-affirms this point:

> […] I think that within all scholarly research there's an understanding between publishers – and authors that the author is using his or her sources in an ethical and professional way. I think there's always been that understanding. If academics started asking questions about whether – whether sources are being abused in any way – then in some respects that calls into question the very nature of an academic's role as a researcher. […] [$K_8$]

Although acting in an ethical and professional manner is the cornerstone of academic best practice, in a significant minority of instances researchers do not act appropriately as was shown by the case of Hwang and others. The scrutiny of research and the underlying data is at the core of a robust higher education system [$D_{21}$]. While 'the channels of distribution have been democratised over the last few years' [$D_8$], as in the previous two cases, the key disseminator of the analysis and interpretation remains the journal publisher [$P_{10}$]. Therefore, where data and/or their metadata are made discoverable on the Web, such as FLLOC and SPLLOC, this is only helping to strengthen checking procedures for publishers and research users [$O_8$].

---

[468] **Mr D.** summarises this point:

> […] I think that research data has a very close relationship to research publications. The – the data is the foundation upon which a number of publications are made. Unless people can look at the data, how can they verify the findings? […] [$D_9$]

[469] Within the academic community, there appears to be a lack of consensus over the usefulness of full verbatim transcriptions, as it is often challenging to match the written word with the tone and context of the spoken word [$K_4$].

#### 6.3.4.6 Data sharing

Academic research data are increasingly seen as one of the most valuable resources arising from research [$M_7$]. In a stark contrast to some areas of chemistry research, there has always been a culture of sharing within second language acquisition research [$O_{33}$]. In the past, second language acquisition researchers would typically rely on data from a very small sample of participants often over a long duration (one to two learners is not unusual although larger studies are also common) [$O_7$].

Sharing high quality data through corpora such as FLLOC and SPLLOC means that other research users can access and re-use data for other purposes [$O_7$], such as for comparison, research in the same fields, and applied research (including pedagogy). **Dr O.** anticipates that the number of corpora and data sharing is set only to increase over the coming years [$O_{35}$]. As shown in the previous two case studies, data sharing over the Web also helps form the basis for future international collaborations; for instance the LANGSNAP research team have worked with partners in France, Mexico and Spain to produce guidelines [$P_{10}$].[470]

FLLOC and SPLLOC are not only used by second language acquisition researchers, but to inform and potentially enhance teaching practices of modern languages teachers and lecturers [$O_5$, $P_{10}$], or by researchers in other disciplines such as phonetics [$O_{36}$]. As **Dr O.** summarises:

> It's very rich data. And, because we've asked certain questions of the data – doesn't mean to say that someone else using the data would use it in the same way. So – and that's the beauty of having the sound files and the – and the transcripts [$O_{36}$].

Moreover as a result of the recent government funding cuts to residences abroad, the UK's University Council of Modern Languages (UCML) is interested in using

[470] **Dr P.** further explains:

> […] For us this [open access] it's important. This is not just about telling other scholars what we've done but – *erm* – teachers – so people who may not have access to journal articles. *Um* – policy makers. […] We see the impacts as being much broader, and also much more *er* international. So for example […] we have project partners in Mexico, in France and in Spain. And […] we work together in producing kind of documents and *er* guidelines and things like this – *er* which are not going to be in academic journals – because our … our aims I think are multiple. […] So the Internet hasn't changed […] the analysis and things that we're doing – but it has changed who we talk to, and who we're going to communicate with. [$P_{10}$]

LANGSNAP to demonstrate the real value of the year abroad (as part of BA Modern Languages) to UK policy makers [$P_{10}$].[471]

FLLOC and SPLLOC further enable non-academics involved in language acquisition (for example teachers and policy makers who do not belong to an institution that subscribes to a number of publishers, or where articles are not open access) to easily gain access to these underlying data and other useful materials that would not be appropriate for academic publication [$P_{10}$].

## 6.4 FLLOC and SPLLOC: interim conclusions

FLLOC and SPLLOC have successfully built sustainable and re-usable corpora of second language acquisition data. All the datasets openly released on the FLLOC and SPLLOC websites are in formats that are open and simple to use, such as XML. Research users therefore are not required to learn or purchase new software.

From the Young Learners Corpus and the LANGSNAP Corpus, it is evident that the projects are subject to rigorous ethics procedures stipulated by the universities involved, the funding bodies, and CHILDES. As minors were involved in the Young Learners Corpus, the data originators had to undergo mandatory CRB checks. In the LANGSNAP study, as many of the undergraduate students were experiencing living abroad for the first time, the data originators involved with the corpus had to undertake counselling training to safeguard the welfare of the participants. The further utilisation of consent forms that gave permission for data re-usability and for open access, the employment of pseudonyms, the removal of (potentially) sensitive and personal information, and independent data checks for confidentiality and quality, all led to the release of ethically sound second language acquisition data.

In consequence, it is clear that FLLOC and SPLLOC largely resolve the third area of concern raised by Hwang and others, by effectively maintaining and disseminating access to data from human participants – many of whom are children – and in the process succeeded in balancing a range of difficult issues around permissions, data protection and confidentiality.

---

[471] *The UK University Council for Modern Languages (UCML) Website* <http://www.ucml.ac.uk/> [accessed 9 August 2015].

However, a potential weak point of both FLLOC and SPLLOC may be the lack of a visible data licence agreement(s). In accordance with the Basic Rules for Data Usage (CHILDES' ground rules); all data made available through CHILDES are governed by an Attribution-NonCommercial-ShareAlike 3.0 Unported licence. The use of a well-known open licence is an advantage, as it makes clear that research users have the permission to re-use the datasets for non-commercial purposes subject to attribution and share-a-like stipulations. However, while FLLOC and SPLLOC state that the datasets are openly re-usable subject to attribution, there is no mention on the FLLOC and SPLLOC websites of the specific Creative Commons licence. Are these statements enough – or should the websites embed the Creative Commons licence information?

There are no known instances of academic misconduct connected with FLLOC and SPLLOC (as with the previous case studies – MEDIN, and eCrystals and LabTrove). FLLOC and SPLLOC's provenance metadata, legal, technological and socio-cultural frameworks are mostly proven as robust, via evaluation of the primary source and interview materials. It is self-evident that through these frameworks FLLOC and SPLLOC confront many of the worst practice ambiguities found in the case of Hwang and others, particularly on ethics issues.

Where data and/or their metadata are made discoverable on the Web, such as FLLOC and SPLLOC, checking procedures immediately become available for publishers and research users. This helps to address the aforementioned weakness in the journal publication process and strengthen the reputation of quality publications and publishers. As with the Hwang case, the interviews revealed that it is not common for academic journals to check or ask for underlying second language acquisition or life histories research data. Again, the default position of the publisher is to trust and have confidence in the data analysis and interpretation undertaken by the researcher. While it must be acknowledged that only a minority of academics will submit articles based on erroneous data, this remains a key weak point of the journal publication process as the essential arbiter of academic validation and scrutiny.

FLLOC and SPLLOC overcome a number of problems that arise through the maintenance of and access to data gathered from human participants, successfully disseminating some of the most difficult data to release as open data. In answer to the question 'what makes for excellent quality academic research?' the same five key themes arise from this case study as in the previous two chapters: (1) accreditation, (2) sustainability, (3) working towards a common understanding, (4) offering a good user

experience, and (5) discoverability. These themes are now explained within this interim conclusion, and will be further evaluated as part of Chapters 7-8.

The robust ethics procedures and independent quality checks the Young Learners Corpus and the LANGSNAP Corpus had to satisfy are testament to the significance of accreditation for sustainable and re-usable academic research data. These ethical assurance processes are a vital part of the socio-cultural and legal frameworks that support the collection, management and (if possible) re-usage of academic research data gathered from human participants. Moreover, diverse approaches are required to cover multiple sensitive scenarios from obtaining data via school children and university students to political dissentients and whistle-blowers.

Sustainability is at the core of FLLOC and SPLLOC. FLLOC is an open data pioneer that has taken advantage of and adapted to technological change. This is clearly demonstrated through FLLOC's initial and continuing commitment to safeguarding its second language acquisition data as a sustainable resource for future re-usage. This is because the data are likely to have a long shelf life even where the child participant is now an adult. For instance, the FLLOC website was established eight years after the start of the first corpus in 1993 to share datasets from a series of related corpora. However, owing to FLLOC's dedication to long-term preservation, the original tape recordings from the initial corpus were digitised and openly released via the website. In addition, FLLOC paved the way for the later creation of its sister website SPLLOC, centred on non-native Spanish learners.

This case study highlights that it is not only web technologies that have greatly impacted on knowledge transfer in a digital age, but the equipment used to both record and store data. For researchers collecting oral data (such as within the domains of second language acquisition and life histories research and for use within this thesis), the continual advancement in recording equipment has significantly enhanced its sound quality and longevity.

Moreover, consent forms are integral to sustainability by providing an essential means of recording mutual understandings between researcher and participant (or their parents/guardians where they are younger than sixteen years old). A strict set of parameters is therefore established to outline how the collected data will be used, published and maintained by the researcher and (potentially) re-used by other research users. Therefore, the data originators obtain rights clearance to openly release these data through the FLLOC and SPLLOC websites both now and in the future. However,

obtaining rights clearance for data collected prior to the Web is often problematic for the re-usage of data, as the capabilities of this technology would not have been accounted for.

As with the previous two chapters, limited funding is raised as a significant issue for sustainability. This is shown by the lack of post-project funding to sustain the FLLOC and SPLLOC websites.

FLLOC and SPLLOC's relationship with CHILDES has helped to work towards a common understanding of second language acquisition data maintenance, interpretation, analysis and dissemination. CHILDES (another open data pioneer) was founded in 1981 to address un-coordinated transcription coding systems. Therefore, it not only provides an international database of first and second language acquisition research data, but CHAT and CLAN as tools for standardised analysis and formatting of transcriptions, along with its mandatory ground rules. With over 2000 published articles (correct in 2003), 4500 subscribers and holding 44 million words (correct in 2007), CHILDES is truly helping to work towards a common understanding for technological, legal and socio-cultural standards in child language acquisition domains on a vast and global scale.

However, CHILDES again demonstrates that one size does not fit all. As a data custodian (such as MEDIN) it manages diverse datasets from child language, bilingual language and second language acquisition data to language disorder data. The latter understandably requires a much stricter permissions process. CHILDES has therefore rightly taken into account the varying degree of sensitivity/protection pertaining to certain types of data, and recognition that not all data can be released as open data. Moreover, a format-driven approach to ethics issues results in some researchers going beyond the mandatory requirements in the pursuit of ticking all the boxes, and some participants may want to partake in research without signing a consent form.

As with the eCrystals and LabTrove case study, there appears to be a limited common understanding of existing legal practice. Yet again, the interview participants were extremely self-reflective about this issue. For instance, while compliance with the Data Protection Act 1998 was cited as significant, the interview participants pondered whether researchers really consider legislative requirements in strict legal terms. Therefore, should ethical and/or legal issues be the principal focus of the researcher or is the higher education institution better-placed to advise researchers of the these issues?

Notwithstanding a lack of enhanced discoverable provenance metadata, the FLLOC and SPLLOC websites provide research users with a good user experience. For instance, the majority of the data are available in XML, which as an open, universal and non-proprietary format. Research users are therefore not required to learn new technical skills or purchase software. CHILDES not only offer technical guidance and global standards (CHAT and CLAN) for transcription, but provide a dedicated team of developers who constantly strive to improve the software for the research community through releasing new software versions and updates. Moreover, the CHILDES developers rapidly respond to user queries through an active email list.

As the FLLOC and SPLLOC websites were developed by iSolutions, the University of Southampton's professional IT service has also proved to be essential to the development of this model. Guidance and support is an essential part of providing a good user experience and sustainability. The FLLOC and SPLLOC researchers also extend their working knowledge of CHILDES to interested academics by running workshops on its advantages and how to use the system effectively. The FLLOC and SPLLOC websites also offer guidance on how to use CHILDES.

Discovery is important for all academic research data re-usage models, because research users need to be able to find the specific data they require with ease. For the FLLOC and SPLLOC data, discovery is facilitated largely through the print context of this research (the published textbook) and the reputation of the researchers involved. In contrast to the case studies in Chapters 4 and 5, FLLOC and SPLLOC do not employ a system of enhanced machine-readable or machine-understandable provenance metadata. This is therefore identified as a feature for potential future improvement. Although the use of machine-understandable provenance metadata could make the second language acquisition data more navigable, this case study acts as a reminder that it is the content and quality of the provenance information which is of principal importance and not whether provenance information is released in machine-understandable format.

In addition to the five key themes raised through the case study analysis, the interviews further highlighted two important grey areas for consideration: (1) the potential negative impacts of the Web on data sharing; and, (2) the integration of data analysis tools within academic data re-usage models.

An important grey area raised in this chapter is the negative impact of the Web regarding data sharing. The instantaneous, global network has sparked a confidentiality and/or privacy paradox. People increasingly share their personal data on the Web, but

also remain concerned about the potential monitoring, profiling and surveillance of their online activities. On one hand, participants may be more reserved (therefore producing more controlled and/or limited data) or individuals may decline an invitation to contribute to research altogether. Whereas, thanks to longstanding use of social media, younger generations with perhaps a more relaxed attitude to privacy may influence future data sharing and research methodologies.

As the amount of academic research data continues to grow so do the opportunities for greater data analysis, such as using text and data mining techniques. It is not just the scientific disciplines employing software solutions to derive more value from datasets; this chapter clearly shows how data analysis plays a big role in second language acquisition research. Through models such as FLLOC and SPLLOC, research users are now able to re-use data from hundreds of language learners leading to more representative findings. In order to do this, there are a number of analysis tools available to researchers. However, the long-established problems of burying underlying and outlying academic research data (confronted by models such as, MEDIN, eCrystals, FLLOC and SPLLOC) could re-surface, as academic research data are potentially concealed under layers of analysis.

Many key themes and grey areas have been raised by Chapters 4-6. Chapter 7 will now re-visit them through cross-comparative analysis and further evaluation to model best practice for excellent quality academic research data re-usage in a digital age, and highlight the principal areas for future research in the concluding chapter.

# Chapter 7: Recommendations

## 7.1 Print to ePrint evaluation

A key motivation for this thesis was to examine the extent in which the re-usage of academic research data has been re-versioned for a digital age. Any recommendations therefore need to be put in context of what has changed with the advent of the Web. Has the model of dissemination – namely the peer-reviewed academic journal – been re-versioned, or has a more fundamental shift taken place?

The academic publication model rightly retains its prominence in a digital age. However, ePublication provides a digital means of disseminating academic research data by offering the same distribution method through a new medium – the Web. **Dr G.** raises this vital point, in Chapter 5, by highlighting how the traditional knowledge transfer process within chemistry has been re-versioned, but not yet re-purposed:

> […] All we've done in chemistry is *er* turn the very conventional process electronic. So all we've done is *um* speeded things up, *er* made things more accessible in terms of the fact that *um* I can download it in seconds […] but effectively I'm getting the same level of information as a PDF as opposed to a photocopy. […] It's just an electronic version of what we've been doing for two hundred years. […]It [the Web] hasn't really impacted much on – on reliability, robustness, fitness for purpose, because we need access […] not to the PDF […] we need access to the actual data – *er* the actual observations, the actual recordings […] in the lab *er* to be able to *er* understand precisely what somebody did; and whether the […] inferences they made, the conclusions they drew, were correct or not. [$G_{16}$]

As the main focus of ePublication continues to be on interpretation and syntheses, many academic publishers still do not acknowledge, let alone address, the need to preserve the underlying and surrounding academic research data on which conference papers, journal articles, monographs and scientific reports are built. For instance, **Dr O.** has never been asked to provide underlying academic research data by an academic publisher [$O_{8,}$ $K_8$].

While some academic research data journals are in existence, they are currently uncommon. They may also fail to capture outlying datasets where these do not fall within core research or where they contain negative results. These journals may also be unable to keep up with the rapid pace of e-data generation.

The mechanism of peer-review remains as essential to research in a digital age as it was within the print era. However, it continues to be unusual for the journal peer-review process to access and independently appraise the underlying academic research

data. Many academic publishers are failing to recognise the value of academic research data as an important entity in their own right. This not only removes a barrier for uncovering potential bad and/or worst practices, but overlooks the building of academic research data corpora that have the potential to further enhance human knowledge and understanding via multiple (often unforeseen) re-uses beyond those purposes for which they are originally designed. This is clearly demonstrated by FLLOC and SPLLOC where the datasets have not only enriched second language acquisition research (where researchers would typically rely on data from a very small sample of participants often over a long duration), but have been re-used by researchers from diverse disciplines from pedagogy [$O_7$] to phonetics [$O_{36}$]. As **Dr O.** summarises:

> It's very rich data. And, because we've asked certain questions of the data – doesn't mean to say that someone else using the data would use it in the same way. So – and that's the beauty of having the sound files and the – and the transcripts [$O_{36}$].

In the case of LANGSNAP, the open release of the data has also led to international collaborations with partners in France, Mexico and Spain to produce guidelines [$P_{10}$].[472] Moreover as a result of the recent government funding cuts to residences abroad, the UK's University Council of Modern Languages (UCML) is interested in using LANGSNAP to demonstrate the real value of the year abroad (as part of the BA Modern Languages) to UK policy makers [$P_{10}$].[473]

In summary, while it is crucial for fundamental longstanding print processes to endure in their (often) re-versioned forms, the ways in which knowledge transfer can be re-purposed, and therefore further enriched, for a digital age should not be overlooked. The three cases studies demonstrated a number of different approaches to storing, safeguarding and disseminating (where appropriate) academic research data on the Web. Excellent quality research must prevail in spite of continuing challenges e.g. with the rise in the number of researchers, the increase of data-driven research, a growth in

---

[472] **Dr P.** further explains:

> […] For us this [open access] it's important. This is not just about telling other scholars what we've done but – *erm* – teachers – so people who may not have access to journal articles. *Um* – policy makers. […] We see the impacts as being much broader, and also much more *er* international. So for example […] we have project partners in Mexico, in France and in Spain. And […] we work together in producing kind of documents and *er* guidelines and things like this – *er* which are not going to be in academic journals – because our … our aims I think are multiple. […] So the internet [sic] hasn't changed […] the analysis and things that we're doing – but it has changed who we talk to, and who we're going to communicate with. [$P_{10}$]

[473] *The UK University Council for Modern Languages (UCML) Website* <http://www.ucml.ac.uk/> [accessed 9 August 2015].

storage capacity, and advances in analysis techniques. These challenges provide an opportunity for those involved with ePublication to consider how such data models can be used to supplement, enrich and overall strengthen the traditional route to dissemination in a digital age as well as going further to preserve valuable outlying data.

## 7.2 Five principles for improved modelling of best practice

Chapters 4-6 focused on academic research data re-usage in very different disciplinary domains from the marine environmental sciences and chemistry to second language acquisition research and life histories. However, the same five fundamental themes emerged in each chapter: (1) sustainability, (2) working towards a common understanding, (3) accreditation, (4) discoverability, and (5) a good user experience. These five fundamental themes are therefore identified by this thesis as five principles for improved modelling of best practice in a digital age.

### 7.2.1 Sustainability: gather data once and use many times

MEDIN, eCrystals, LabTrove, FLLOC and SPLLOC all identify sustainability as a principal component for modelling best practice. This is fittingly summed up by the principle: gather data once and use many times, which is highlighted by both **Ms. E** and **Mr. N** in Chapter 4. While this principle is exclusively identified by the MEDIN case study, it equally applied to eCrystals, LabTrove, FLLOC and SPLLOC, and other academic data re-usage models.

All case studies confronted significant problems with data collection and re-use where similar data are gathered multiple times and/or not shared. For instance, within the marine environmental sciences, data are generally expensive to collect and often require specialised equipment, field work and vessel hire. Before MEDIN was founded in 2008, there was significant duplication of surveys which constituted a waste of vital resources and time.

For many highly collaborative disciplines the continual turnover of staff and/or more limited engagement with a research project from some parties due to PhD involvement, grants and other commitments, can undermine the sustainability of data. eCrystals and LabTrove were therefore established in response to these continuity issues to not only make data re-usable for the research community but those continuing and

new staff directly involved with the collection, maintenance and dissemination of academic research data.

Ongoing maintenance is vital for securing the long-term preservation of academic research data. However, the level of funding required to continually manage their legal, technological, socio-cultural and provenance metadata frameworks may not be readily available. In consequence, budget constraints can threaten their sustainability. The issue of funding is further raised in section 8.4.2 (within Chapter 8) as a key grey area for future work.

The potential to openly release data instantly to a global audience via the Web also has a negative impact on the future collection of certain types of data. For instance, some people are more reticent about participating in research due to concerns over the accessibility of sound recordings and transcripts. This potential sustainability issue was raised during Chapter 6 and is also highlighted in section 8.4.7 (within Chapter 8) as a key grey area for future work.

### 7.2.2 Working towards a common understanding

MEDIN, eCrystals, LabTrove, FLLOC and SPLLOC all demonstrate that best practice for academic research data re-usage is most effective when a user-community both embraces and commits to agreed, and very high, standards. No model encapsulates this principle of community involvement better than MEDIN, which thrives on bringing together interested parties through working groups to discuss, influence and adopt common standards of best practice – that seek to be more than just the sum of different parts:

> […] You need to get all the – all the relevant users and holders of the data that you're dealing with *um* round the table. [...] Try to build that consensus is – is [an] incredibly difficult task and one that MEDIN have done very well. […] There's a danger in these sort of partnerships that people sort of say: 'yeah, yeah, it's brilliant, it's great', but don't actually […] commit to it. But I think that's something that comes with time, because eventually you build a critical mass […] of organisations, and data, and resources – that if you're not involved in it you're actually losing out. […] [$E_{27}$]

CHILDES similarly has helped FLLOC and SPLLOC to work towards a common understanding for technological, legal and ethical standards. Furthermore, eCrystals and LabTrove are also directly engaged with the open laboratory movement and working towards a common understanding by using open standards and open source software.

Encouraging people to change their existing best practices is one of the most significant barriers to working towards a common understanding. People need to be shown the benefits of better practice regarding renewed means of data analysis and synthesis. Extra effort and in some cases cost is required to alter formats, provenance metadata standards and switch to a different publication process. For some models such as MEDIN that inherit a diverse, and historic, collection of data and data management practices, adapting to change can be a more lengthy process. Then again these models also benefit from an established breadth of expertise that more recent models do not necessarily encompass.

It is important then that standardisation is an organic process accepted by a user-community. However, top down data policies from national, European and (in some cases) international funding councils and legislative bodies can also facilitate greater harmonisation. One such example is the INSPIRE directive. However, it is unclear to what extent such new regulatory standards would help change existing data sharing practices in other disciplinary domains, where significant political drive to standardise data re-usage is currently lacking.

While models should strive to move as close as possible to a shared approach to academic research data re-usage, one size will never fit all. As clearly demonstrated by Chapters 4-6, data are a diverse entity with each dataset emanating from a specific context. In consequence rights holders have different re-usage preferences. For instance, CHILDES manages diverse datasets including language disorder data, which requires a much stricter permissions process than many other datasets it holds. Therefore, not all data can be released as open data, because of the varying degree of sensitivity/protection pertaining to certain types of data. Improved modelling of best practice for academic research data re-usage therefore takes into account the need for a shared approach with inbuilt flexibility.

Consensus on the long-term preservation of a particular dataset can be challenging, as much effort is required for its safeguard. However, premature deletion of a dataset may impede a future (academic) discovery, which had the potential to be unlocked through a new approach, such as the future application of a new algorithm or piece of laboratory equipment. The measurement of the current and potential future value pertaining to a specific dataset is an extremely difficult task. As a result, value metrics is presented as a key grey area for future research in section 8.4.9 (within Chapter 8).

### 7.2.3 Accreditation

Robust quality checks are important to MEDIN, eCrystals, FLLOC and SPLLOC not just through ethics boards and funding bodies, but by domain specialists. While it is difficult to fully understand whether accreditation is an essential feature of LabTrove, as a researcher can set up their electronic laboratory notebook with the purpose of releasing data only where it has been approved by a supervisor or colleague, accreditation is still an important aspect.

Data originators need to undertake quality checks on the data collected alongside other independent assessments. The seven MEDIN accredited thematic data archive centres meet twelve minimum standards of best practice, and therefore conduct independent quality assurance procedures on any datasets which are received.

For eCrystals, FLLOC and SPLLOC quality assurance largely occurs at the project or laboratory level. For instance, the research assistant who transcribed a particular sound recording as part of the Young Learners Corpus would not assure the quality of that resulting transcription. As is the case with eCrystals and signing off a crystal structure for deposit, another member of the team would be responsible for those checks. Due to the formulaic nature of crystallographic data, free online software is also used to automatically examine a dataset's integrity and consistency. This provides the research user with an additional quality report and means of checking the research produced.

To further strengthen quality, eCrystals employs a traffic light system to signal different levels of data quality to research users: red alert – there are some serious problems with the data that require explanation; amber alert – there are potential problems with the data; and, green alert – there are potential areas for concern. This is an effective way of warning research users of different quality levels. Quality data should not be confused with perfect data, as **Mr C.** explains:

> […] The best thing to do is to get into […] an area culturally where everyone recognises that nothing is perfect. So, there's a big difference between something being fraudulent *er* which clearly it's quite right that this kind of approach deters people from fraudulent practice, because they have to make a certain amount of their working available *um* to a community which is happy to correct each other's data […] another institution may spot that you've made a few mistakes in your data, but then – you know, the reverse will happen and in fact the benefit to the sector is that the data is much cleaner – and everybody has a much better quality based data to be working on. *Um* and that's the kind of

> collaborative approach to data cleaning which is very positive, but quite hard to achieve in a competitive environment, but I think […] it can be achieved. [$C_8$]

For instance, while the Young Learners Corpus has been extensively quality assured by the project team a minor number of post-project data corrections have been made by the research team to rectify errors or mistakes previously missed. Similarly in crystallography, it is extremely rare to obtain perfect data with no quality alerts therefore even the green alert carries a data warning. In addition, data warnings enable researchers to share negative results, which can prevent unnecessary duplication of a failed methodology and/or experiment.

While provenance metadata have a vital role in the re-usability of academic research data, the amount of metadata needs to be manageable for effective accreditation. It is likely that that the next generation of laboratory equipment will increase the production of computer-generated provenance metadata (see Chapter 5, section 5.3.1.1). The result will be a likely increase in the amount of provenance metadata; further necessitating the production of meta-metadata (a record of how the provenance metadata were captured). Vast amounts of provenance metadata and meta-metadata could therefore potentially jeopardise the re-usability of academic research data, as greater effort, time and resources will be required for their effective generation and management. Moreover, academic research data with extensive provenance metadata and meta-metadata may be seen as less re-usable, as research users may feel overloaded with information. This meta-metadata issue is therefore raised in section 8.4.5 (within Chapter 8) as a key grey area for future work.

In addition to examining issues concerning best practice in the authorised management of academic research data, there needs to be better-understanding of mismanagement. For instance, unauthorised released of academic research data could also have negative ramifications (see Chapter 5), particularly with commercial, personal and/or sensitive datasets where a data leak could prove to be a contravention of data protection law, a trade secret or harm a patent application. For instance, an unauthorised release of highly personal and sensitive life histories data could threaten a research participant's personal safety. Mismanagement of data often has more immediate consequences in the digital age where a small number of clicks may lie between a closed data record and its instantaneous release to a global audience via the Web. Unauthorised releases of data are thus highlighted in section 8.4.8 (within Chapter 8) as another key grey area for future work.

### 7.2.4 Discoverability

In order to facilitate re-usage, research users need to be able to locate where academic research data are stored. Not all academic research data are digitally accessible and may be stored offline, with no web delivery available. In addition, not all data are catalogued in print archives. However, a signpost needs to be in place just to flag their existence (where possible), because not all data will be released at the point of access.

Provenance metadata are an essential requirement for signposting the place in which data is released. Particular key words can be used to enhance discovery through search engines. Online discovery is an obvious advantage over print era discovery mechanisms. Nonetheless, while extremely useful in the digital age, search engines are not the only mode of discoverability. For instance, in the case of FLLOC, the reputation of the authors and associated publications still constitute an important discovery mechanism.

Provenance metadata are further required to both draw together related datasets and distinguish between similar datasets. MEDIN employs its discovery metadata to assemble associated marine environmental data that are dispersed over a number of thematic data archive centres, which were gathered under the same theme, discipline, particular project and/or survey. As **Ms E.** further explains:

> […] The whole concept of MEDIN, this idea of distributed data archive centres [and] a – a central metadata portal – without metadata it wouldn't hang together. […] You couldn't link a metadata record to a – a dataset, or link several datasets together where they're collected under the same survey. […] some datasets we collect have *er* a sort of chemical component, a biological component, geology component – and with the metadata it allows us to store these within these thematic centres of excellence, but also they can be aggregated through the metadata. […] [$E_{17}$]

For research users re-using crystallographic data, the ability to access individual datasets rather than in pre-defined families of structures is beneficial. Provenance metadata therefore enable researchers to personally identify relationships between crystal structures, which are useful for collective analysis.

A definitive version of an academic research dataset has different disciplinary meanings. For FLLOC and SPLLOC, the corpora are offered as definitive versions. However, for other disciplines such as crystallography, provenance metadata are also used to distinguish between several versions of a dataset, as **Miss J.** states:

> […] publishing several different versions of the same dataset – *um* that actually to my mind that is a benefit not a problem […] there are circumstances under

> which sometimes you will process a dataset to the best of your ability *um* and […] you might publish it, as I say, an incomplete dataset. […] and then you process it again later using a different piece of software, or with more experience behind you, or any number of circumstances. And when you do that, you produce a better or a different result. Now actually you might want to publish both results as an illustration of the fact that you're new software is more advanced in its ability to refine the data structure. […] I don't see that kind of data duplication as a problem particularly if it's stored with metadata that allows you to disambiguate the two entries. […] [$J_{19}$]

**Miss J.** further explains how provenance metadata are able to prevent confusion where datasets are intentionally duplicated:

> […] publishing the same crystal structure gathered in two different locations twice, *um* that is a problem, but […] if the proper automated systems were put in place to check for data duplication, then I foresee that will not necessarily be a problem. And again it could be a benefit in that someone using a – a different difractometer they're on a different side of the world – there's always that idea of scientific validation by repetition from a different, disassociated research group. [$J_{19}$]

FLLOC and SPLLOC impart a new perspective on provenance metadata not found in the previous two case study chapters (Chapters 4-5). The success of academic research data re-usage does not necessarily lie in the sophistication of the technology that underpins the provenance metadata. Although it is often assumed the more complex the IT aspects of academic research the more searchable and re-usable the academic research data becomes, it is the quality of the metadata that is of primary concern. Although machine-understandable provenance metadata would aid discovery (for example via search engines) and more succinct categorisation (such as the metadata fields used in the case studies of the earlier chapters) would offer a more concise way to attain the key metadata, this is a secondary concern. Therefore, FLLOC and SPLLOC demonstrate that the creation of robust metadata comes first; their organisation and enhanced machine-readability (or machine-understandablity), although important, are a secondary consideration.

Signposting and (in some instances) releasing academic research data through data platforms are essential aspects of online data discovery. However, there are some instances where multiple data platforms are attempting to signpost and release data. For instance, the bathymetric survey data held by the UKHO are (potentially) caught between five discovery portals (see Chapter 4, section 4.3.4.3). As a result, multiple data platforms with overlapping remits can seriously impede the discoverability of academic research data and compromise the principal of gather once and use many

times. This multiplicity issue is therefore raised in section 8.4.6 (within Chapter 8) as a key grey area for future work.

### 7.2.5 Good user experience

It is clear from Chapters 4-6 that research users require models that are both easy to use and offer a high level of multiple utility. Ultimately researchers do not want to have to purchase new software or learn new technical skills:

> […] I think the critical thing is to make the information and the data format itself as simple as possible. So that people *um* don't have to learn new software […] or buy new software in order to access datasets. I think the – the data files themselves need to contain as much information as possible in a way that's easy for people to just kind of click and open. […] [$S_3$]

As MEDIN, eCrystals, LabTrove, FLLOC and SPLLOC all release data under open formats, there is no requirement for research users to purchase or learn new software. However, the use of open formats is not enough; models have to provide a sufficient level of functionality. For example, the website needs to be easy to navigate and research users expect to locate any relevant provenance metadata without delay (in the fewest clicks possible). In consequence, those models perceived as requiring too much effort to learn without adding high levels of value are likely to drastically limit their user base.

Chapters 4-6 provide an insight into different levels of user experience. In the case of MEDIN, its open community approach to creating its IT infrastructure and commitment to simplicity is testament to its success. Despite the read-me style provenance metadata, the FLLOC and SPLLOC websites are also easy to navigate. In contrast to this, eCrystals provides an example of a more limited user experience. It has more restrictive search functionality than other chemical inventory systems, because its IT infrastructure – EPrints – was designed for the archiving of print publications rather than academic research data. The management and re-usage of crystal structures substantially differs from that of publications. Its lack of functionality contributes to its limited user base of five known users. However, those directly involved with eCrystals are aware of the shortfalls of its IT infrastructure and have improvement plans in place subject to securing the necessary funding.

Another fundamental component of a good user experience is a high level of support and guidance. MEDIN offers significant support to its user community through its discovery metadata tool and dedicated helpline. Moreover, its thematic data archive

centres do not require provenance metadata to be compliant with MEDIN discovery metadata standards on submission. The added benefit for MEDIN is its five member core team, dedicated to overseeing its daily operation and access to data specialists at its accredited data archive centres. For smaller laboratory-based models such as eCrystals and LabTrove and project based models such as FLLOC and SPLLOC, the provision of a data management team is just not a possibility because of budget limitations. In the case of eCrystals, LabTrove, FLLOC and SPLLOC, data management is at user-level.

Support and guidance is not only required by research users, but those involved with depositing and managing academic research data. For instance, eCrystals ensures that it trains new team members about data quality and management to avoid the release of poor quality data through eCrystals. For the researchers involved with the Young Learners Corpus and the LANGSNAP Corpus, support comes from the CHILDES developers who rapidly respond to user queries through an active email list. Furthermore, the FLLOC and SPLLOC researchers also extend their working knowledge of CHILDES to interested academics by running workshops on its advantages and how to use the system effectively.

For researchers who are considering whether to set up an academic research data re-usage facility, but perhaps do not have the technical expertise required, it is important to factor in technological assistance within any funding bid. In the case of FLLOC and SPLLOC, iSolutions (the University of Southampton's professional IT service) created the websites and databases.

Next generation academic data re-usage models are likely to further enrich the user's experience through the provision of greater data analysis functionality. However, there is the potential for these widespread tools to re-bury academic research data under computer-generated analysis. This issue of computer-generated analysis is therefore highlighted in section 8.4.1 (within Chapter 8) as a key area for future work.

Despite access to legal experts (for instance as part of a higher education institution's dedicated legal services), it is clear that researchers have limited legal awareness. As academic research data re-usage becomes more complex (e.g. through the use of data and text mining), the legal issues involved may also become more complicated and require greater legal guidance. A crucial example of this legal complexity is the management of licensing and attribution stacking; a key grey area for future work outlined in section 8.4.4 (within Chapter 8). In consequence, more work needs to done to explore how to increase legal awareness within the academic

community. Legal guidance tools are therefore emphasised as a further significant grey area for future work in section 8.4.3 (within Chapter 8).

## 7.3 Recommendations summary

It is imperative for all those involved with the generation and management of academic research to consider how to support and guarantee the entire lifecycle of an academic research dataset and/or resource. Without an agreed and applied process for sustainability, an academic research dataset will fail to achieve the objective of gather data once and use many times. An improved model for best practice starts with this pre-determined strategy for sustainability by working towards a common understanding of the required provenance metadata, legal, technological and socio-cultural frameworks. These robust frameworks are required to inform high standards for the collection, management and re-usage of academic research data from the outset. Such high standards may already exist (e.g. through organisations similar to CHILDES), or may require active development (e.g. via working groups such as those utilised by MEDIN).

Accreditation through robust (independent) quality checks is needed throughout the lifecycle of a dataset to safeguard its integrity and therefore ongoing re-usability. It is imperative to identify legal and ethical requirements, codes of practice and standards from authoritative organisations (e.g. institutional, funding and standards bodies) and appropriate licensing. Academic research data that have been collected, managed and accredited to high standards are then ready for release or signpost (where this is possible). A high level of discoverability is achieved through high quality provenance metadata that can be used not only to signpost data accessibility, authorisation and quality, but can link across multiple datasets. On discovering these academic research data, the data platform (point of access) needs to offer a good user experience. At the most basic levels of required functionality, research users want a system that is easy to navigate where they can locate any relevant provenance metadata without delay – in the fewest clicks possible – and there is no prerequisite to purchase or lean new software.

As a final point of recommendation, these five principles for improved best practice in a digital age should not be an afterthought or add-on, but embedded throughout the lifecycle of an academic research dataset and/or resource for effective future re-usage. This thesis now turns to Chapter 8 which outlines its conclusions and nine key grey areas for future work.

# Chapter 8: Conclusion and future work

## 8.1 Conclusion and future work: introduction

This thesis began by asking: what makes for excellent quality academic research data re-usage in a digital age? It further explored how best practice should be modelled now and in the future by examining what could be learnt from longstanding practices. Modelling best practice for academic research data re-usage in a digital age is a complex and multi-faceted challenge despite the wealth of research protocols such as: long-established peer-review through academic publishers and funding bodies; research councils' policies; digital libraries; open licensing; ethics boards; IT systems and search engines; charitable organisations such as the Committee on Publication Ethics (COPE); academic forums to debate the validity of research; and individuals actively recording instances of academic misconduct such as Retraction Watch, Embargo Watch and Copy Shake and Paste.[474]

However, as the motivating case of Hwang and others demonstrates, these research protocols still do not prevent concerted academic misconduct or the wilful publication of erroneous academic research and data in some instances. Although an apparently self-evident fact – the case studies in Chapters 4-6 reinforce that best practice is only as good as the researcher undertaking it. In the final months leading up to the completion of this thesis retractions and spoof articles in a number of key journals continued to surface, as indicative of only a small proportion of inaccurate research. On 13 February 2014, Paul Jump reported in the *Times Higher Education Online* that an article published in 2006 by *Cell Metabolism* and authored by a number of researchers at Oxford University was retracted due to image manipulation.[475] Furthermore, the Top

---

[474] *Committee on Publication Ethics (COPE) Website* <http://publicationethics.org/> [accessed 9 August 2015]; Adam Marcus and Ivan Oransky, 'Retraction Watch Blog', *Retraction Watch Website* <http://retractionwatch.wordpress.com/> [accessed 9 August 2015]; Ivan Oransky, 'Embargo Watch Blog', *Embargo Watch Website* <https://embargowatch.wordpress.com/> [accessed 9 August 2015]; Debora Weber-Wulff, 'Copy, Shake and Paste: A Blog about Plagiarism and Scientific Misconduct', *Copy, Shake and Paste Website* <http://copy-shake-paste.blogspot.co.uk/> [accessed 9 August 2015].

[475] Paul Jump, 'Former member's misconduct causes third retraction for lab', *The Times Higher Education*, 13 February 2014 <http://www.timeshighereducation.co.uk/news/former-members-misconduct-causes-third-retraction-for-lab/2011262.article> [accessed 9 August 2015].

10 Retractions of 2014 published by the co-founders of the Retraction Watch Blog – Adam Marcus and Ivan Oransky – in *The Scientist* also proves an interesting read.[476] Finally, as highlighted by Chapter 2, the controversy around stem cell research dominated by the Hwang case saw two retracted articles by Haruko Obokata and others published in *Nature* on 2 July 2014.[477]

Examples of worst practice research in major peer-reviewed journals were underscored by a number of spoof articles published in this period that were generated by computer programs, not researchers. On 4 October 2013, John Bohannon published an article in *Science* claiming that his spoof article was accepted for publication by no fewer than 157 academic journals.[478] These included academic journals published by such authoritative organisations as Elsevier, Wolters Kluwer and Sage.[479] However, the article was also rejected by 98 academic journals.[480] On 24 February 2014, Richard Van Noorden reports in *Nature News Online* that the prestigious academic journal publishers Springer and IEEE had withdrawn over 120 nonsensical, computer generated papers.[481]

As was emphasised in Chapter 2, spoof knowledge claims are based on fabricated academic research data, but they do not have the intention to mislead, even defraud, research users. They often seek to test the very system that assures the reliability of academic research data for research users. As a feature of print culture, spoof knowledge claims are not a new phenomenon. What is new is the potential for artificial intelligence to create papers able to test the main peer review channels, such as academic journals and conferences. As a result researchers do not have to spend time

---

[476] Adam Marcus and Ivan Oransky, 'The Top 10 Retractions of 2014', *The Scientist*, 23 December 2014 <http://www.the-scientist.com/?articles.view/articleNo/41777/title/The-Top-10-Retractions-of-2014/> [accessed 9 August 2015].

[477] Haruko Obokata and others, 'Retraction: Bidirectional developmental potential in reprogrammed cells with acquired pluripotency', *Nature*, 511 (7507) (2014), 112 <http://dx.doi.org/10.1038/nature13599>; 'STAP retracted: Two retractions highlight long-standing issues of trust and sloppiness that must be addressed', *Nature*, 511 (7507), (2 July 2014), 5-6 <http://dx.doi.org/10.1038/511005b>

[478] John Bohannon, 'Who's Afraid of Peer Review?' *Science*, 342 (6154) (4 October 2013), 60-65 <http://dx.doi.org/10.1126/science.342.6154.60>

[479] *Ibid.*

[480] *Ibid.*

[481] Richard Van Noorden, 'Publishers withdraw more than 120 gibberish papers', *Nature News*, 24 February 2014 <http://dx.doi.org/10.1038/nature.2014.14763>

manually producing spoof articles.[482] However, an excess of spoof articles could also offset its benefits by exerting too much pressure on peer-reviewers who are already managing increased volumes of article submissions.

In consequence, while the answer to the primary question: 'what makes for excellent quality academic research data re-usage in a digital age?' may seem self-evident; the response to this question is that it lies in the many diverse and robust research protocols and people safeguarding research in a digital age. However, these research protocols are not enough. This thesis has therefore sought to critically examine key areas for concern and recommended five principles for improved modelling of best practice for academic research data re-usage in Chapter 7.

## 8.2 Methodological approach

While Chapter 2 made clear that critical examination of modelling best practice for academic research data re-usage is not new and found within a number of disciplines, this thesis offers a new synthesis of existing research by joining up and filling the gaps between previously unconnected areas in ways not hitherto achieved. In most instances, existing research was shown as: too theoretical, trapped within disciplinary silos, too policy-driven, and/or lacking historical context. For these reasons, the key literature is often scattered within and across a number of disciplinary domains, therefore making it more disjointed. In constructing a literature review for this thesis, it was very difficult to determine which were the most influential texts, and which should take principal position within a literature review crossing interdisciplinary domains.

Moreover, most of the research appears to be caught within a 'rhetoric of newness', a point emphasised through Lauren Rabinovitz and Abraham Geil's consideration of digital culture.[483] As a result, significant emphasis is often placed

---

[482] Ian Sample, 'How computer-generated fake papers are flooding academia', *Guardian*, 26 February 2014 <http://www.theguardian.com/technology/shortcuts/2014/feb/26/how-computer-generated-fake-papers-flooding-academia> [accessed 9 August 2015].

[483] Lauren Rabinovitz and Abraham Geil state:

> Discussions about digital culture assume that new computerized technologies provide such fundamental rupture from the past that there are no continuities or, worse, that they willfully obliterate the past in creating new models. Such ahistoricism is problematic be-cause it tends to reproduce at the level of scholarship what is one of the hallmarks of digital culture – its rhetoric of newness. It is painfully obvious that this is neither the first technological revolution in human

primarily on the digital open data and open access debates without first addressing how academic research data have been, and therefore need now to be, prepared for effective and longer term re-usage.

The Web has enabled data sharing to flourish through its instantaneous speed of transfer, its vast and increasing storage capacity, its potential global reach, and ease of search (engines; copy, cut and paste). However, research users now have increased expectations concerning academic research data re-usage in a digital age. In theory, academic research data are no longer confined by print and restrained by physical delivery; data have the potential to be immediate, mutable and reach wider audiences than ever before. In consequence, open movements (such as open access, open data, open laboratories, open licensing, and open source) have gained greater momentum and influence in a digital age.

This thesis shows that these open movements are not new digital phenomena, for example the concept of the republic of letters was voiced by Thomas Bodley to facilitate a community for knowledge transfer in the 1600s.[484] Therefore, this thesis addresses who and what has made the open release of academic research data possible where it was unattainable in the past. For this reason, the main focus of this thesis is on how to safeguard the re-usage of academic research data both now and in the future.

The debates about more academic data becoming increasingly open are only one consideration. If there has been one key finding in this thesis, it is that open data are only as effective as the quality of a producer of a dataset and its provenance, and the trust and confidence the research user places on a particular dataset. For instance, academic research data which are openly released may not have been readily quality checked, whereas those in an obscure format may have been. In order for re-usage to be most effective, academic research data need to be continually guaranteed as reliable, robust and fit for purpose, but also be disseminated in standard formats.

---

> history nor an event independent from its cultural heritages and historical roots, and so a rhetoric of newness is at best a myopic one.

Lauren Rabinovitz and Abraham Geil, 'Introduction' in *Memory Bytes: History, Technology, and Digital Culture*, ed. by Lauren Rabinovitz and Abraham Geil (Durham NC: Duke University Press, 2004), pp.1-2. Google eBook.

[484] G.R, Evans, 'Academic Libraries and the Law: What Legal Protections Guarantee the Survival of Britain's Academic Library Collections?' *Education Law*, 4 (2008), 248 (p. 248); Paul Keen, *The Crisis of Literature in the 1790s: Print Culture and the Public Sphere* (Cambridge: Cambridge University Press, 1999), p. 4.

By firmly placing this thesis within interdisciplinary web science practices, which specifically consider the issues arising across three research domains – the digital humanities, law, and web and internet science – this thesis is better-positioned to understand the four key interconnected frameworks required for ensuring excellent academic research data generation and hence re-usage in a digital age. As a key finding across all the cases studies in Chapters 4-6, these are: provenance metadata, legal, technological and socio-cultural frameworks.

### 8.2.1 Case studies

By using the case of Hwang and others (in Chapter 2) as a motivating example of worst academic practice, this thesis was able to rapidly highlight some of the main loopholes in the longstanding peer-reviewed academic publication model, which assures good academic data usage and re-usage in a digital age. The key issue for this thesis, therefore, was how to avoid interpretations built on erroneous datasets and their dissemination (failures of citation were not the issue) in order to circumvent bad and worst practice where possible.

Through the utilisation of Hwang and others as an initial and motivating case study, this thesis was able to accomplish more than just responding to key readings. It was able to better-engage and advance different parts of the web science field. It recognised that there was not only an intellectual need to model best practice, but a pragmatic one. In consequence, it identified five best practice principles (in Chapter 7) that are workable across disciplines within the sciences, social sciences and humanities, and recognise the diverse nature of academic research data. These five best practice principles therefore seek to prevent and immobilise minor instances of bad practice to major incidents of academic misconduct.

In consequence, the thesis examined three case studies (across the sciences, social sciences and humanities) in UK Russell Group universities to raise consciousness of the complexities around managing and releasing diverse types of academic research data on the Web, regardless of discipline or research methodology. These case studies were used to supplement the literature review where much of the research was within disciplinary silos and therefore had not joined together the four key provenance metadata, legal, technological and socio-cultural frameworks for academic research data re-usage. Moreover, the case study approach proved an effective and efficient method of

probing a wealth of information available in the secondary literature, but which often lie outside the traditional academic publication route. Again, the fact that these three case studies deliberately lay outside the thesis author's prior higher education experience was crucial to prevent pre-given assumptions concerning academic research data re-usage.

To further enrich the case studies by going beyond a primary and secondary literature review, semi-structured interviews were conducted with eighteen participants (six per case study). To cover the wide range of provenance metadata, legal, technological and socio-cultural issues raised by each case study, the participants selected had wide-ranging roles. To aid cross-comparison and maintain consistency, each case study covered the following five areas of expertise: (1) data management, (2) data policy, (3) academia, (4) legal and (5) technology.

A highly detailed and distinct outline of the methodological approach taken is not a necessity for some disciplines, such as the humanities and law. However, this thesis is written for an interdisciplinary audience in order to find commonalities regarding research data re-usage in a digital age. In consequence, Chapter 3 is a vital component of this thesis, as it explains, justifies and outlines the chosen methodological approach. It further acts as a preamble to Chapter 6 which also focuses on the University of Southampton's best practice for the collection, maintenance and (re)use of academic research data gathered from human participants.

Chapter 4 then focused on how diverse types of academic research data from multiple originators, contexts and sources can be safeguarded for a wide set of research users. By using MEDIN, which guarantees data across the marine sciences and is made up from a five member core team and seven accredited thematic data archive centres, this case study was able to demonstrate how a robust core framework of discovery metadata enabled maintenance and re-usage of data on such a wide-ranging and large scale.

Chapter 5 explored how the maintenance of and access to academic research data generated by multiple originators within changeable collaborative research teams can balance a range of difficult problems with mutual permissions, sustainability, communal quality control and joint authorship. By utilising eCrystals, this case study was able to show how crystallographic researchers at the University of Southampton were able to openly release data generated through their experiments, but where the samples had been gathered by researchers at other institutions. The electronic laboratory

notebook LabTrove added further value to this case study by providing a different platform in which to share data generated via experiments.

Chapter 6 turned to second language acquisition and life histories research to address how maintenance of and access to sensitive and personalised data can balance a range of difficult problems concerning permissions, data protection and confidentiality. By focusing on FLLOC and SPLLOC, this case study shows how these researchers are able to openly release data normally considered among the most difficult to disseminate because they were gathered from minors and young adults. A focus on life histories research further enriches this case study by providing access to **Dr K.** who has experience of collecting, managing, using and re-using data that are not just personal but highly sensitive. Life histories researchers sometimes encounter extremely confidential data gathered from adult participants, who may be political dissidents or whistle-blowers. These data are principally collected through interviews, memoirs and other personal materials. If there were any breach of anonymity leading to the uncovering of their true identities, this could have a major impact on their personal safety and that of their families, their employment, and/or their reputation.

## 8.3 Statement of conclusions

Peer-review remains a cornerstone of academic best practice for research generation, analysis and re-usage including in academic journals. Retaining the cumulative value of peer-review in the digital age and in these e-print mechanisms is vital. However, knowledge transfer has yet to be fully re-purposed for the digital age. For instance, some researchers still have to distil the details of an experiment into a set number of pages as was the case with print media. However, underlying and supporting academic research data no longer need to be hidden. MEDIN, eCrystals, LabTrove, FLLOC and SPLLOC are all examples where knowledge transfer is being re-purposed to encourage the long-term preservation of and (where possible) access to academic research data. These academic research data are considered as highly valuable in their own right. Given that analytic techniques such as text and data mining are opening up new possibilities for re-usage not widely available in an age of print, these data platforms will provide key resources of high quality academic research data.

Provenance metadata will continue to have an increasingly central and pivotal role in sustainable, discoverable and re-usable resources of accredited academic

research data across the sciences, the social sciences and the humanities. Academic research data and provenance metadata are likely to be more automated and interoperable, as a result helping to facilitate the principle: gather once and use many times. This greater automation raises a number of grey areas for future research. These will need to be confronted to ensure that modelling best practice continues to evolve, in order to confront these mounting challenges of automation and interoperability on both a national and global scale (see sections 8.4.1 and 8.4.5).

Legal rights management and clearance is a fundamental part of capitalising on the growing knowledge economy. However, legal guidance seems to be the least resourced area – open licensing solutions are not enough. Individuals do not only need understanding of intellectual property rights issues, but (as has been shown by the case studies) other areas of law including data protection and legally-enforced metadata standards. Therefore, this thesis raises two significant legal grey areas for future work: the need for more legal guidance tools (see section 8.4.3); and, how the academic community can respond to increasing licensing and attribution stacking (see section 8.4.4).

Through an understanding of a longer knowledge transfer heritage, this thesis has shown the concepts of data accessibility, trust, authorisation and traceability still stand in the digital age as they did in the pre-digital age. These principles necessitate robust provenance metadata, legal, technological and socio-cultural frameworks. The five key principles – (1) sustainability, (2) working towards a common understanding, (3) accreditation, (4) discoverability, and (5) a good user experience – raised by this thesis in Chapter 7 are indispensable to achieving these robust frameworks and therefore effective academic research data re-usage.

Thus far digital media and processes seem to be just bigger and faster ways of enabling knowledge transfer. Linked data offers a more intelligent and integrated vision of academic research data re-usage. However, to what extent has the Web been truly transformative and therefore revolutionary in terms of knowledge transfer? This is the key philosophical question for future web scientists to contemplate.

## 8.4 Future work

This thesis concludes by outlining nine key grey areas for future work, as revealed by the three case studies in Chapters 4-6. In order to continue to re-purpose and strengthen academic research data re-usage in a digital age better-understanding is therefore required of the following: (1) computer-generated analysis, (2) funding, (3) legal guidance tools, (4) licensing and attribution stacking, (5) meta-metadata, (6) multiplicity, (7) negative impacts of the Web, (8) unauthorised releases of data, and (9) value metrics. These nine key grey areas for future work are now explained in sections 8.4.1-8.4.9.

### 8.4.1 Computer-generated analysis

Next generation academic data re-usage models will not only be providing data but data analysis tools. The provision of greater data analysis functionality is raised as the third and final grey area within the FLLOC and SPLLOC case study. Chapter 6 clearly demonstrates how data analysis software not only plays a big role within scientific disciplines, but across the humanities. A number of data analysis tools are already available to second language acquisition researchers. Moreover, a text and data mining exception is being introduced into copyright law to enable researchers to text and data mine for non-commercial purposes.[485]

While it is not disputed that wider access to data analysis tools would bring greater benefits for the advancement of knowledge, it is no panacea and requires careful handling. This thesis has clearly demonstrated how data re-usage models such as MEDIN, eCrystals, LabTrove, FLLOC and SPLLOC confront the long-established problems of human-generated analysis that often bury underlying and outlying academic research data. However, computer-generated analysis potentially risks re-burying academic research data through configurations of derived data. The key issue here is ensuring that the computer-generated analysis is not separated from the analysed

[485] 'Exceptions to Copyright: Research', UK Intellectual Property Office Document (October 2014), *UK Government Website* <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/315014/copyright-guidance-research.pdf> [accessed 9 August 2015].

datasets. Provenance metadata are vital to ensure that the reliability of such analysis can be traced back to high quality datasets.

Maintaining the links between computer-generated analysis and its underlying datasets requires serious input from researchers with technological expertise, particularly in the realm of artificial intelligence. For instance, how can systems be built with high levels of traceability to help ensure that the underlying data can be automatically exposed, and each computer-generated analysis has a comprehensive audit trail?

### 8.4.2 Funding

The sustainability of an academic research data model in a digital age has the potential to be compromised by limited funding, which in turn threatens the integrity of its IT infrastructure and quality of its datasets. Budget constraints were not only raised as a grey area through the MEDIN case study, but also recognised as a significant issue within the two subsequent case studies. For instance, while MEDIN were aware of the benefits of linked data, there was limited funding to implement this (at the time of the interviews). Moreover, due to government funding constraints MEDIN was unable to advertise the model to attract a wider pool of users. As budgets vary dramatically and are dependent on the specific type and circumstances of data acquisition, storage and dissemination – how can economic factors be better managed to reduce the risks to sustainability?

A lack of funding to support new technological advancements (e.g. new technical standards and enhanced query and analysis functions) not only poses risk to the future integrity of the IT infrastructure, but potentially to the quality of the academic research data in question. As with many projects across the sciences, the social sciences and the humanities there is not often the funding available to actively maintain the data and database after project completion. For instance, while post-project data corrections were carried out on the datasets collected by the Young-Learners Corpus, these appeared to be on a voluntary basis.

This thesis can only raise awareness of this funding issue and call for those involved with data maintenance to consider the cost of management not only in the short term, but also for future preservation. Clearly a funding strategy and an institutional support strategy are required, but how it should be shaped lies outside this

thesis. This grey area requires an intervention from research institutions, funding councils and policy makers (principally from the Department for Business, Innovation and Skills (BIS)) who have the necessary expertise to apply an economics approach.

### 8.4.3 Legal guidance tools

Given Chapters 4-6 all show that more needs to be done to work towards a common understanding of legal rights as researcher awareness of these is very limited, this is raised as an overarching grey area (it does not appear as an explicit area for future work within any of the case studies' interim conclusions). None of the models are legal entities, and all rely on support from legal professionals. However, as with many legal service departments within higher education institutions, there is not the capacity to advise each researcher about the release and re-use of every dataset. Researchers may confuse legal rights, and choose unsuitable licences. Legal professionals have an important role regarding knowledge transfer practice.

Chapters 4-6 all raised the need for a degree of flexibility. If best practice is seen as too prescriptive it can weaken data gathering and re-usage rather than add value, such as a format-driven ethics approach. While the value of consent forms is not disputed, some researchers go beyond the mandatory requirements in the pursuit of ticking all the boxes. As a result, excessive restrictions may prohibit future re-usage.

More research needs to be conducted into how to best support researchers. For instance, to what extent would the academic community benefit from more on-demand guidance such as online case studies? How could legal professionals have a greater role in ensuring the authorised re-usage of data?

However, legal guidance tools should only be provided as a support for and not a replacement to legal experts. Such tools may help to facilitate dialogue between researchers and legal experts, and help researchers to consider legal issues before academic research data are released rather than as an afterthought.

### 8.4.4 Licensing and attribution stacking

As a result of text and data mining, a new dataset may be derived from hundreds of datasets from multiple sources and of mixed quality and provenance. The provision of a robust provenance metadata framework that captures all these data is not an easy task.

Furthermore, each dataset may be subject to a different data licence agreement leading to licensing stacking – a grey area for future research highlighted in Chapter 5:

> […] With research data […] as soon as you start merging data from multiple sources your data is subject to all of the restrictions of all of the sources. And is subject to all the requirements of all the sources, and *um* the more different licences, the more painful […] [$R_{15}$]

Rights clearance and management are more significant issues for larger data custodians such as MEDIN. As MEDIN largely safeguards third party data, it needs to ensure that it has the authority to store and signpost each dataset. Moreover, the thematic data archive centres all operate different licensing frameworks. This issue was raised by the MEDIN case study – how can data custodians, such as MEDIN, better manage this licensing complexity?

On the surface, the solution to licensing stacking is simple – the utilisation of one standard licence agreement where all academic research data are subject to the exact same terms and conditions. Due to the diversity of data originators, sources and contexts a one-size-fits-all approach is not possible however. As was shown by the literature review in Chapter 2, alongside organisation-specific licences, there are also a number of open licences to choose from. For instance, Creative Commons offers six standard licences.

Despite some initiatives to offer one standard licence, such as the Open Government Licence as one standard licence for open government data, such solutions only partially solve the issues of licensing stacking. Often government bodies will openly release data under the Open Government Licence whilst also attaching an organisation-specific attribution statement. Within the academic community, many higher education institutions and funding bodies favour the Creative Commons by Attribution Licence (CCBY). Although this may resolve licensing stacking issues, it does not address the problem of attribution stacking, as **Mr C.** explains:

> […] the new RCUK policy, *um* on open access for publications, has specified the use of a CCBY licence. And, they're saying […] one of the reasons behind that is that they want to really maximise the re-use of those data, including *um* harvesting by text mining […] which I think is great in principle, but in practice you're still not getting 'round *er* the need for attribution. A lot of licences still say attribution is a requirement and […] I do think what you get […] attribution stacking. You get so many people that you need to attribute when you're doing this text mining – that in fact it becomes impossible very, very quickly to fulfil the attribution part of it. So you either need to look at a licence where you become, you know, not worried about the attribution – which a lot of people are concerned about, because you know then you're divorcing yourself from the

> kind of credit for the work as it were, and I fully you know understand that. *Um* but then if you hold onto the attribution as being key […] how do you get […] around this issue? So I think there's still a lot to do in that area. [$C_{11}$]

Therefore, models managing data from multiple originators, contexts and sources and research users mining large quantities of these data must be able to handle licence and attribution stacking. To complicate matters further, the international nature of knowledge transfer through the Web means that these issues not only have to be dealt with on a national level but must confront multi-party collaborations and multi-jurisdictional licensing. This is an issue raised in both Chapters 5 and 6. More research is required to address the issues of licensing and attribution stacking, especially from the legal informatics community.

### 8.4.5 Meta-metadata

Chapter 5 raised the issue that academic research data, their provenance metadata and analysis are likely to become increasingly computer-generated. As a result, endless chains of meta-metadata could be generated, which is to the detriment of the quality of research findings. The MEDIN case study showed that the simple principle of getting 'everyone in the same room' [$B_{22}$] may help to resolve this question, because best practice would be informed by expert practitioners.

### 8.4.6 Multiplicity

Model multiplicity can not only jeopardise the sustainability of a model, but adversely impact on the discoverability of academic research data. This was raised a grey area in Chapter 4; where bathymetry surveys data are already experiencing this through five different portals attempting to do the same task. However, access point multiplicity is also occurring across other disciplines. For instance, at one end of the scale, datasets may be made available for re-use through the website connected to specific laboratory (such as is the case with eCrystals in Chapter 5) or as part of collection of related projects (such as is the case with FLLOC and SPLLOC in Chapter 6). At the other end of the scale, data may be deposited at institutional level within a repository (such as Soton ePrints), a national repository (such as the UK Data Archive) or an international repository (such as CHILDES).

Model multiplicity can seriously compromise the principal: gather once and use many times. Perhaps, the best way forward is to release provenance metadata and if

appropriate a dataset once, but enable other platforms to harvest and/or link to this definitive version (where possible).

### 8.4.7 Negative impacts of the Web

The Web has sparked a confidentiality and/or privacy paradox. It is made clear by the first grey area raised by Chapter 6 that the impacts of the Web on knowledge transfer are not all positive. On one hand, participants may be more reserved (therefore producing more controlled and/or limited data) or individuals may decline an invitation to contribute to research altogether. However, this position may change as younger generations – with perhaps a more relaxed attitude to privacy – begin to influence future data sharing and research methodologies. For instance, how will this impact on areas of research such as life histories where open access has long been at its core? It would be valuable for this area to be taken forward by social science researchers – to gain a better-understanding of how future academic data re-usage might be altered by these negative impacts.

### 8.4.8 Unauthorised releases of data

Some unauthorised releases of data have negative ramifications, particularly with commercial, personal and/or sensitive datasets where a data leak could prove to be a contravention of data protection law, a trade secret or harm a patent application as **Miss J.** states:

> Patents will always have an impact. […] They'll have […] a two-fold impact from my understanding in that *erm* – you've got the first thing of well people don't really want that data releasing full stop – if they can have it. *Er* they're going to want it embargoed up until the date in which the patent is published at the very least – is the other effect. *Um* because of course, if we have an accidental release of data – something goes wrong – there's a malfunction in the system – whatever. *Er* then obviously that voids the patent, because information has been released in the public domain before the patent has been put in place. [$J_8$]

To what extent unauthorised releases of data are more likely to be caused by a system malfunction or human error is an area requiring further investigation. More research needs to examine the risks involved with knowledge transfer. To what extent do those involved with the collection, management and re-usage of an academic research dataset need to consider and implement security measures? What academic research data re-usage platforms are the most secure and appropriate for a particular dataset?

### 8.4.9 Value metrics

Long-term preservation of data requires a considerable amount of time and effort in order to verify, annotate, curate and cleanse data. The key questions to be considered are: how do researchers and their institutions quantitatively and qualitatively determine the value of each dataset? How do they decide where to invest resources and time into their maintenance? Quantitative user metrics are not enough to settle on a value, as a dataset that does not seem highly valuable at the time of collection could prove to be invaluable in different future re-use, for example when a new algorithm or piece of laboratory equipment is applied. Moreover, the preservation of the definitive version of dataset alone may not be enough to safeguard future re-usage. Data originators and institutions need to consider whether to retain additional information, such as past versions of a dataset and their provenance, and even store a virtual machine where necessary.

More research is urgently required to address the issue of long-term preservation, and ensure that valuable data are not lost because of a misguided decision on the extent of their usefulness. It would also be of further interest to examine how and to what extent past data have recently been re-used and re-discovered as invaluable to new knowledge transfer.

### 8.4.10 Summary: future work

It is crucial these nine grey areas for future work are taken forward by web scientists and other researchers across a variety of disciplines, including law, data science, computer science, digital humanities, economics and sociology. This thesis has demonstrated that in order to properly engage with improved modelling of best practice, there first needs to be cognisance of the longer heritage of knowledge transfer and its four vital provenance metadata, legal, technological and socio-cultural frameworks. Therefore, it is further imperative that future research confronts ahistoricism and appreciates the interdisciplinary context of this area.

As a final point, it is of paramount importance that the five best practice principles – (1) sustainability, (2) working towards a common understanding, (3) accreditation, (4) discoverability, and (5) a good user experience – should facilitate both intellectual and pragmatic approaches to high quality academic research data re-usage both now and in the maturing digital age.

# Appendices

# Appendix A Ethics documentation

## A.1 University of Southampton: Ethics review checklist (2011)

UNIVERSITY OF Southampton
School of Law

**Summary of key ethical guidelines for research**

Researchers should inform participants about the aims of the research, likely publication of findings in the context in which they will be reported and of potential consequences for participants. Participants should provide informed consent, before participating in the research, and be made aware that they are free to withdraw at any time.

Researchers should respect their participant's confidentiality. Researchers should protect the information that could identify individual participants, and ensure security of their data. For example, researchers are advised to use pseudonyms for participants, store data securely, and dispose of confidential data carefully (for example, shredding).

Researchers should communicate their findings and the practical significance of their research in clear, appropriate language to relevant research populations and other agreed stakeholders. Researchers should avoid fabrication, falsification or misrepresentation of evidence, data and findings. Plagiarism should be avoided. Authors are obliged to cite sources of information or ideas drawn from elsewhere.

All staff/students should be aware of the University's ethics policy:
http://www.southampton.ac.uk/inf/ethics_policy.html

## University of Southampton School of Law: Ethics Review Checklist (Student)

This ethics checklist must be completed by the student **before** commencing work on the dissertation, and **handed into the PGR Programme Director**.

**Please Tick (□) one**: ~~Undergraduate □~~ ~~Postgraduate (Taught) □~~ **MPhil/PhD** ☑

**Degree programme/Certificate: MPhil/PhD Web Science**

| | | | |
|---|---|---|---|
| **Your Name:** | **Laura German** | **ID number:** | [redacted] |
| **Univ of Soton Email:** | leg406@soton.ac.uk | | |
| **Supervisor:** | **Professor Saxby (Law), Dr Carr (Computer Science) and Professor Orr (Humanities)** | | |

| |
|---|
| **Dissertation Title:**<br>**Academic Research Data, Provenance and Copyright Law** |
| **Expected start date and duration: October 2010 – 3 years** |
| **Funder (if applicable): Web Science Doctoral Training Centre (EPSRC)** |

| **Part 1**<br>Does your research involve any of the following? | **YES** | **NO** |
|---|---|---|
| 1. Interviews | Yes | - |
| 2. Questionnaires/Surveys ***(a draft of the questionnaire/survey must be attached to this form)*** | - | No |
| 3. Analysis of personal or corporate details (e.g. bank records, personnel or admin records, test results etc.) that are not already in the public domain (e.g. published in a book) | - | No |
| 4. Participant observations | - | No |

***If you have answered 'NO' to all of the above then your research does not need any further ethical consideration. Please sign and date on page 2, see your supervisor for their approval and signature and then hand this form into the PGR Programme Director.***

***If you answered 'YES' to any question then please continue on to Parts 2-4 below.***

| **Part 2** | YES | NO |
|---|---|---|
| 1. Does the study involve participants who are particularly vulnerable or unable to give informed consent? (eg children, adults with special difficulties etc) | - | No |
| 2. Will the study require the co-operation of an advocate for initial access to the groups or individuals? (eg children, people with disabilities, adults with a dementia etc) | - | No |
| 3. Could the research induce psychological stress or anxiety, cause harm or have negative consequences for the participants (beyond the risks encountered in their normal lifestyles)? | - | No |
| 4. Will deception of participants be necessary during the study? (eg covert objectives or observation of people)? | - | No |
| 5. Will the study involve discussion of topics which the participants would find sensitive (eg sexual activity, drug use)? | - | No |
| 6. Will financial inducements (other than reasonable expenses or compensation for time) be offered to participants? | - | No |
| 7. Are there problems with participants' right to remain anonymous, or to have the information they give not identifiable as theirs? | - | No |

| | | | |
|---|---|---|---|
| 8. Is there any way the participants might be unaware of their right to freely withdraw from the study at any time? | | - | No |
| 9. Will the study involve recruitment of patients or staff through the NHS? | | - | No |
| 10. Does the study involve any sort of confidential data that may need to be destroyed at the end of the study? | | Yes | - |
| Please indicate the anticipated number of study participants | Adults:10-15 | Minors: None (under 18) | |

**Part 3** For each item answered 'YES' in Part Two, please give a summary of the issue and action to be taken to address it.

*Does the study involve any sort of confidential data that may need to be destroyed at the end of the study?*

The interviews, subject to the participants' consent, will be recorded by the use of a Dictaphone. All of these recordings will be deleted when the interview has been transcribed and the MPhil/PhD is completed.

The consent forms signed by interview participants will be destroyed at the end of study – when the MPhil/PhD is completed.

*Please continue on a separate sheet if necessary*

~~**Part 4** If you answered yes to question 3 in Part 1in respect of corporate information, has approval been received from the organisation for the use of their name, data or access to employees? YES/NO~~

~~*If yes, please attach evidence, such as a letter from the company, and indicate below the nature of the evidence.*~~
~~*If no, explain how confidentiality issues will be addressed.*~~

**Signatures: This Section must be completed**

**Signed** (Student) ..... Date: .....

**Signed** (Supervisor) ..... Date: .....

***Please hand this form into the PGR Programme Director. Incomplete forms will be returned.***

***For office use only:***
Formed received by: Date:

**To be completed by the designated member of the School Ethics Committee:**

☐ Appropriate action taken to maintain ethical standards – no further action necessary

☐ The issues require the guidance of the School's Ethics Committee

***COMMENTS:***

Signed: ..... Date: .....

*Notified to School Ethics Committee* : ..... *(Date)*

*Notified to Research Governance Office*: ..... *(Date)*

## A.2 Semi-structured interviews: participation information document (January 2012)

*Participants have the unconditional right to withdraw from the study at any time and for any reason.*
*Participants may take away this 'Participation Information Document' for their future reference.*

# Semi-Structured Interviews
# 'Participation Information Document'

**January 2012**

**Preliminary Research Title:**
'Copyright is not enough. How can more appropriate and effective academic research data reusability be better facilitated on the Web?'


**Investigator:**

**Laura German**
*MPhil/PhD Web Science Candidate*

Member of the Web Science Doctoral Training Centre,
Faculty of Business and Law,
University of Southampton,
SO17 1BJ
leg406@soton.ac.uk

Faculty of Business and Law, University of Southampton

*'Academic Research Data, Provenance and Copyright Law Project.'*

**Approved by the Management Ethics Committee August 2011 (further additions approved in January 2012.) – Paper System.**

Please refer to the 'Ethics Clearance Documentation' for further information.

*Participants have the unconditional right to withdraw from the study at any time and for any reason.*
*Participants may take away this 'Participation Information Document' for their future reference.*

## Contents

*Participants have the unconditional right to withdraw from the study at any time and for any reason.*
*Participants may take away this 'Participation Information Document' for their future reference.*

## 1. SEMI-STRUCTURED INTERVIEW PROCEDURE

### 1.1 Before the Interview

*Researcher Information:* Participants will be provided with brief biographic information about the researcher and the aim of the research.

*Interview Information:* Participants will be emailed electronic copies of the following (all part of the 'Participation Information Document'):

- Semi Structured Interview Procedure
- Consent Information
- Consent Form

The researcher will also bring physical copies of these documents to the interview for the participants to keep as future reference and to sign.

*Prior Reading of Questions:* At the beginning of the interview, participants will be given two minutes to read through the brief list of semi-structured questions that will be asked.

### 1.2 During the Interview

*Duration:* Interviews will normally last up to one hour.

*Consent Information:* Before the interview begins, interview participants will be given a paper copy of the consent information to keep for future reference. Participants will be (given a couple of minutes to read through this document and) asked if they have read the information (sent by email in advance) and if they have any questions about this information (that may arise).

*Consent Form:* Once (if) participants have asked and had answered satisfactorily any questions that have arisen (and are satisfied), they will be asked to fill in the consent form - initialling the statements, signing and dating it. The researcher will then sign and date the consent form.

*Ethics Clearance:* Participants will be given a paper copy of the 'Ethics Clearance Documentation' to keep, if they wish, for future information.

*Sound Recordings:* Interviews will be recorded by a Dictaphone. Participants' names will not appear in the sound recordings. The interview will begin with "this is interview *n*."

### 1.3 After the Interview

*Follow-up Questions:* Participants will be asked if they wish to have a copy of the recorded interview (this will solely be for their own private use), and if they would be willing to answer further written questions (by email) the researcher may have at later stages of the project. The same confidentiality will govern any further exchanges between researcher and participant.

*Participants have the unconditional right to withdraw from the study at any time and for any reason.*
*Participants may take away this 'Participation Information Document' for their future reference.*

*Withdrawal Procedure:* Participants have the unconditional right to withdraw from the study at any time and for any reason. Please email Laura German (email address: leg406@soton.ac.uk) to inform the researcher of your official withdrawal from this research. The researcher also retains the right to withdraw from the interview.

*Notice:* In the thesis it will be made clear that: "It must be noted that all views are the participants'; they may not represent the views of [Case Study One –Organisation], [Case Study Two –Organisation], [Case Study Three –Organisation] or the University of Southampton. Any inferences drawn from these interviews belong to the Thesis Author and may not necessarily be held by participants or their respective organisations."

*Future Use:* All of the information collected will be used only for the purposes of this study.

**Thank you for taking the time to help with this research.**

*Participants have the unconditional right to withdraw from the study at any time and for any reason.*
*Participants may take away this 'Participation Information Document' for their future reference.*

## 2. CONSENT INFORMATION

**Preliminary Research Title:**
'Copyright is not enough. How can more appropriate and effective academic research data reusability be better facilitated on the Web?'

**Investigator:**
**Laura German**
*MPhil/PhD Web Science Candidate*

Member of the Web Science Doctoral Training Centre,
Faculty of Business and Law,
University of Southampton,
SO17 1BJ
leg406@soton.ac.uk

### 2.1 Brief Overview:

Laura German is second year MPhil/PhD Web Science candidate at the University of Southampton. Laura's background is in law, she obtained an LLB (Hons) in 2009, but she also has an MSc (Dist.,) in Web Science. Laura's research is interdisciplinary - she has three joint supervisors - Professor Saxby (Law), Dr Carr (Computer Science) and Professor Orr (Humanities).

The aim of this research is to explore the ways in which academic research data reusability can be made more appropriate and effective on the Web. It is assumed that current copyright law is not enough to facilitate this. The objective is to use three illustrative case studies from different academic disciplines, across the sciences and humanities, to explore the shortfalls of copyright and how factors such as provenance metadata can be used to better facilitate academic research data reusability on the Web.

### 2.2 Purpose of Conducting Semi-Structured Interviews:

In order to ascertain how academic data reusability can be made more effective and appropriate on the Web and whether copyright is not enough to facilitate this, it is vital that both theoretical and practical data about this issue is obtained. Therefore, the researcher seeks to conduct semi-structured interviews with participants connected to each illustrative case study who have different roles within the scholarly communication of academic research data. This will give a broad overview of whether copyright does hinder academic data reusability, as rights can conflict, and incorporate the varying experiences of individuals involved in the creation, management and usage of academic research data, from academics to information scientists.

*Participants have the unconditional right to withdraw from the study at any time and for any reason.*
*Participants may take away this 'Participation Information Document' for their future reference.*

### 2.3 Best Practice

**2.3.1 Data to be collected:**

- The participants' names will be obtained upon signing consent forms;
- Their roles within the academic publishing model;
- Answers given in semi-structured interviews (the questions are attached to this 'Ethics Application') will be recorded by a Dictaphone; and
- All of this information will be used only for the purposes of this study.

### 2.4 Processes for ensuring data security

**2.4.1 The separation of identifying data and the anonymisation process/the method of linking the consent form (if any) to the participant's data**

- Participants' names will not appear in the thesis;
- Participants' names will not appear in the recordings;
- A separate list will be created which links participants' names and corresponding anonymous pseudonyms, such as: Joe Bloggs – Mr.Y;
- Therefore, participant Joe Bloggs would be referred to as Mr. Y within the thesis or recording;
- The participants' names will appear only on the consent forms and a list which links participants' names and corresponding anonymous pseudonyms;
- Gender will be assigned at random – to avoid a situation where only one woman is interviewed and ten men – possibly making it more difficult to ensure confidentiality. For example: Mr Smith may be referred to as Miss W; and
- All consent forms and the list, which links participants' names and corresponding anonymous pseudonyms, will be locked in the researcher's cabinet.

**2.4.2 The processes for destruction of data**

- When the thesis has been submitted and awarded final marks consent forms and the list, which links participants' names and corresponding anonymous pseudonyms, will be destroyed by the Faculty of Business and Law, where they are held securely in the researcher's cabinet;
- Therefore, after this time 'identifying data' will cease to exist; and
- It will only be retained for as long as necessary.

**2.4.3 E-Thesis and Publishing**

The University of Southampton mandates the electronic submission of an e-thesis that is openly accessible on the Web.[1] This must be made clear and agreed to by the interview participants. The researcher may also wish to publish the thesis in the future.

[1] http://www.soton.ac.uk/library/research/etheses/students.html [accessed 23 January 2012].

*Participants have the unconditional right to withdraw from the study at any time and for any reason.*
*Participants may take away this 'Participation Information Document' for their future reference.*

**2.4.4 Availability of Sound Recordings**

- The sound recordings will not be made available on the Web; and
- They will only be made available to specific individuals directly involved in the research project – Laura German, supervisors and examiners.

**2.4.5 Transcriptions and Quotes**

- Interviews will not be transcribed;
- The researcher will use important quotes from the interview or show that a number of participants agree/disagree on a certain point;
- The interviews will be recorded and stored on a Dictaphone;
- They will also be stored on the researcher's computer as a back-up file (this computer is password protected);
- The interviews will be burnt onto a CD;
- This will be available for Laura German (the interviewer), Laura's supervisors (Professor Saxby, Professor Orr and Dr Carr) and Laura's thesis examiners to listen to only;
- This CD will only be kept as long as is necessary; and
- All sound recordings will be deleted after final marks are awarded.

**2.5 Subject data is accessible to the participants**

The authority which will give participants access to their subject data will be the Project Supervisors.

**2.5.1 Project Supervisor Contact Information**

*Dr. Les Carr* (Computer Science)
Email: lac@ecs.soton.ac.uk
Web Page: http://www.ecs.soton.ac.uk/people/lac
Faculty of Physical and Applied Sciences,
University of Southampton,
SO17 1BJ
United Kingdom
Telephone Number: +44 (0)23 8059 4479
Fax: +44 (0)23 80592865

*Professor Mary Orr* (Modern Languages)
Email: M.M.Orr@soton.ac.uk
Web Page: http://www.soton.ac.uk/ml/about/staff/mmorr.page
Faculty of Humanities,
University of Southampton,
SO17 1BJ
United Kingdom
Telephone Number: (023) 8059 3408
Facsimile: (023) 8059 3288

*Professor Stephen Saxby* (Law)
Email: s.j.saxby@soton.ac.uk
Web Page: http://www.soton.ac.uk/law/about/staff/sjs.page

*Participants have the unconditional right to withdraw from the study at any time and for any reason.*
*Participants may take away this 'Participation Information Document' for their future reference.*

Faculty of Business and Law,
University of Southampton,
SO17 1BJ
United Kingdom

*Participants have the unconditional right to withdraw from the study at any time and for any reason.*
*Participants may take away this 'Participation Information Document' for their future reference.*

## 3. CONSENT FORM

### 3.1 DECLARATIONS

*Please initial the following boxes if you agree with the corresponding statements:*

1. The participant has read the 'Semi Structured Interview Procedure' and has opportunity to ask further questions about the interview. ☐
2. The participant has read the 'Consent Information' and has opportunity to ask further questions about the study. ☐
3. The participant has received satisfactory answers to any questions posed and any additional information they have requested. ☐
4. The participant understands how personal data provided will be managed and what will happen to it at the end of the study. ☐
5. The participant is made aware that the University of Southampton mandates that all its theses are made openly accessible in its repository on the Web. ☐
6. The participant is made aware that the researcher may wish to publish this research in the future. ☐
7. The participant is made aware that the interview will be recorded using a Dictaphone and burnt onto a CD to be accessed only be the researcher, the named supervisors and examiners. ☐
8. The participant would not mind to be contacted in future, via email, if any further related questions arise. ☐
9. The participant acknowledges that they have the unconditional right to withdraw from the study at any time and for any reason. ☐
10. The participant is made aware that these interviews have been reviewed and approved by the Faculty of Business and Law's Ethics Committee at the University of Southampton. ☐

*Participants have the unconditional right to withdraw from the study at any time and for any reason.*
*Participants may take away this 'Participation Information Document' for their future reference.*

**3.2 SIGNATURES**

**I agree to participate in this semi-structured interview.**

Name of participant: (Please print using block capitals)

.....

Signature of participant:

.....

Date:

.....

Name of researcher:

LAURA GERMAN

Signature of researcher:

.....

Date:

.....

Faculty of Business and Law, University of Southampton

*'Academic Research Data, Provenance and Copyright Law Project.'*

**Approved by the Management Ethics Committee August 2011 (further additions approved in January 2012.) – Paper System.**

Please refer to the 'Ethics Clearance Documentation' for further information.

# Appendix B Semi-structured interview tables

[Ch3-S3.4.5-T2] refers to Table 2: Table of designed semi-structured interview questions located in Chapter 3, section 3.4.5.

## B.1 Chapter 3: MEDIN case study

### B.1.1 Mr B. record of interview questions

| **Mr B.** Record of Interview Questions | | | |
|---|---|---|---|
| **Interview Section:** | **Question Duration:** | **Question:** | **Directly Referenced in thesis?** |
| $B_1$ | 00:00:00 to 00:01:08 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $B_2$ | 00.01.08 to 00.04.33 | **2) i) [Ch3-S3.4.5-T2]** | Yes |
| $B_3$ | 00.04.33 to 00.06.48 | **2) ii)** [**LG** adds: 'I heard *um* when we were discussing before [the interview began] *um* ten years ago there was different standards and things like that.'] **[Ch3-S3.4.5-T2]** | Yes |
| $B_4$ | 00.06.48 to 00.08.08 | **4) [Ch3-S3.4.5-T2]** | Yes |
| $B_5$ | 00.08.08 to 00.09.20 | 'Do you get academics coming to you [MEDIN] who say have published their dataset with a journal and then will say: … 'I'll – they'll publish it with MEDIN as well'?' | Yes |
| $B_6$ | 00.09.20 to 00.11.34 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $B_7$ | 00.11.34 to 00:13:16 | '*Um* so what do you think the way forward for licensing would be? Just having one standard licence?' | Yes |
| $B_8$ | 00:13:16 to 00.15.19 | '*Um* what about where data access is restricted, are there any policies for the future say in the next generation, in a hundred years' time, should there be an embargo period that should be lifted, or has that been thought of?' | Yes |
| $B_9$ | 00.15.19 to 00.16.04 | **6) [Ch3-S3.4.5-T2]** | Yes |
| $B_{10}$ | 00:16:04:16:52 | *'Um* do you think legal restrictions should form part of provenance metadata? – So you [research users] explicitly know boundaries of [re-]use and it's all …?' | No |
| $B_{11}$ | 00:16:52 to 00:17:28 | '*Um* with data, do you [MEDIN] find that you have the creators of the data – but they might need to be shown as the creator in the provenance metadata – but the IPR [intellectual property rights] might belong to – I don't k[now] – their University or something? Does that ever come up where you have to distinguish between the two?' | No |
| $B_{12}$ | 00:17:28 to 00:20:15 | **7)** […] [**Mr B.** asks: 'I'm not quite sure what you mean by ethical issues? **LG** responds: '*Um* – sort of, I don't know, if – well, I think it sort of relates to what we were talking [about] before – habits being exploited, or I think I've read about *um* certain species that are endangered – making sure you don't show, perhaps, where they are and things like that?' **[Ch3-** | Yes |

| | | | |
|---|---|---|---|
| | | **S3.4.5-T2]** | |
| $B_{13}$ | 00:20:15 to 00:20:51 | '*Um* at MEDIN, how much data is open to – available to everyone on the Web and how much is maybe just restricted to people who are scientists working in that field?' | No |
| $B_{14}$ | 00:20:51 to 00:22:04 | 'But do you think for MEDIN well – MEDIN as a portal – do you think for *um* data to be effective and appropriate, well appropriately re-used – that you need it open to everyone or do you think it's better that you keep it to people within a community who understand all about where it's come from and – do you think it's of interest to the wider public?' | No |
| $B_{15}$ | 00:22:04 to 00:23:55 | 'What do you think hinders *um* data re-usability the most at MEDIN? Do you think it's *um* technological issues, legal issues, social issues – or do you think it's social issues around *um* understanding the legal issues?' | Yes |
| $B_{16}$ | 00:23:55 to 00:25:11 | '*Um* do you think the Open Government Licence and initiatives like *um* open government data and Creative Commons have – had an impact on MEDIN?' | Yes |
| $B_{17}$ | 00:25:11 to 00:27:13 | '*Um* in a more general sense, how do you think the Web itself has impacted on putting out marine data? Do you think it's had a considerable impact?' | Yes |
| $B_{18}$ | 00:27:13 to 00:30:00 | **8) [Ch3-S3.4.5-T2]** | Yes |
| $B_{19}$ | 00:30:00 to 00:30:33 | 'Also keeping up with different standards in the years to come [issue for MEDIN in the future]?' | Yes |
| $B_{20}$ | 00:30:33 to 00:31:08 | 'And vocabularies of marine data terms, when you have lists, they'll – they won't change really, very much either?' | No |
| $B_{21}$ | 00:31:08 to 00:32:53 | **9) MEDIN) [Ch3-S3.4.5-T2]** | Yes |
| $B_{22}$ | 00.32.53 to 00.35.00 | '*Um* as a portal – do you think other academic disciplines, for example chemistry, *um* could learn anything from how you're managing data?' | Yes |
| $B_{23}$ | 00.35.00 to 00.36.36 | '*Um* I've – I've found some other sort of *um* countries that have similar marine portals – they're trying to make – Australia and the USA. Have you [MEDIN] sort of worked with them – or do they learn things from you, and you learn things from them?' | Yes |
| $B_{24}$ | 00.36.36 to 00.37.50 | '*Um* sorry one thing I forgot to ask, *um* earlier connected to ethical issues and sensitive data, was if you collect data from outside the UK? […] Do you collect data from outside UK waters?' | No |
| $B_{25}$ | 00.37.50 to 00.39.07 | **10) [Ch3-S3.4.5-T2]** | Yes |
| $B_{26}$ | 00.39.07 to 00.39.59 | 'So, would you prefer to have some sort of MEDIN licence rather than using Creative Commons?' | Yes |
| $B_{27}$ | 00.39.59 to 00:40:06 | '*Um* is there anything else?' | No |
| **Question 3) not asked as this was added after the MEDIN case study interviews took place. See Chapter 3: Methodology, section 3.4.5 for further information.** | | | |
| *Interview completed at 00:40:06* | | | |

### B.1.2 Ms E. record of interview questions

| **Ms E.** Record of Interview Questions | | | |
|---|---|---|---|
| **Interview Section:** | **Question Duration:** | **Question:** | **Directly Referenced in thesis?** |
| $E_1$ | 00:00:00 to 00:01:13 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $E_2$ | 00:01:13 to 00:02:20 | **2) i) [Ch3-S3.4.5-T2]** | Yes |
| $E_3$ | 00:02:20 to 00:03:11 | '*Um* do you think MEDIN, as a portal, is really useful for you [DASSH]? Has it sort of helped you standardise things with other data centres?' | Yes |
| $E_4$ | 00:03:11 to 00:05:36 | **4) [Ch3-S3.4.5-T2]** | Yes |
| $E_5$ | 00:05:36 to 00:06:34 | 'Do you think with MEDIN, if it was publicised more – academics knew more about it – maybe they would change their attitude [towards sharing their data]'? | No |
| $E_6$ | 00:06:34 to 00:08:20 | **5) i) [Ch3-S3.4.5-T2]** | Yes |
| $E_7$ | 00:08:20 to 00:09:12 | 'How much of it [the biodiversity data] is archived with – maybe it's got embargos on it'? | Yes |
| $E_8$ | 00:09:12 to 00:11:06 | '*Um* what about licensing restrictions over datasets? Do they pose big problems?' | No |
| $E_9$ | 00:11:06 to 00:12:51 | 'So do you think there's a lot of room for MEDIN to expand [into] lots of different areas?' | Yes |
| $E_{10}$ | 00:12:51 to 00:14:43 | 'And do you have a lot of say as a data centre in what MEDIN – in the direction of MEDIN?' | Yes |
| $E_{11}$ | 00:14:43 to 00:15:59 | '*Um* I wanted to ask you about – I was reading on the [DASSH] website that you create a lot of maps and charts and things, is that right? With IPR that's not a big issue for you for other people coming along and making new products out of them, sort of mashing them together or …?' | No |
| $E_{12}$ | 00:15:59 to 00:17:11 | **2) ii) [Ch3-S3.4.5-T2]** | Yes |
| $E_{13}$ | 00:17:11 to 00:18:06 | '*Um* would you say the Web's had a big impact on this [data sharing]?' | Yes |
| $E_{14}$ | 00:18:06 to 00:19:34 | 'Do you think open data, such as open government data and that push – that sort of political push at the moment, that's had an impact on you [DASSH]?' | No |
| $E_{15}$ | 00:19:34 to 00:21:48 | **5) ii) [Ch3-S3.4.5-T2]** | Yes |
| $E_{16}$ | 00:21:48 to 00:22:24 | 'Are a lot of people afraid of this [misuse of data]? Is that why maybe academics don't want to give over data as well?' | Yes |
| $E_{17}$ | 00:22:24 to 00:24:22 | **6) [Ch3-S3.4.5-T2]** | Yes |
| $E_{18}$ | 00:24:22 to 00:24:57 | 'And how important is authorial assertion? Saying this – that you know this is the person who created this dataset, and being able to – even if you, say, write about a dataset or use it you can reference that person?' | No |
| $E_{19}$ | 00:24:57 to 00:28:39 | **7) [Ch3-S3.4.5-T2]** | Yes |
| $E_{20}$ | 00:28:39 to 00:30:05 | 'What's your [DASSH's] quality assurance policy?' | No |

| | | | |
|---|---|---|---|
| $E_{21}$ | 00:30:05 to 00:30:33 | **8) [Ch3-S3.4.5-T2]** | No |
| $E_{22}$ | 00:30:33 to 00:31:27 | 'What influenced them [the INSPIRE Directive and the Marine Strategy Framework Directive] in the first place?' | Yes |
| $E_{23}$ | 00:31:27 to 00:33:09 | 'Is [the harmonisation of European Union member states collecting and sharing marine environmental data] that because, you think of the maturity of technology – now you've got the metadata standards and *um* …?' | Yes |
| $E_{24}$ | 00:33:09 to 00:34:32 | 'Have you [DASSH] got any plans for linked data?' | Yes |
| $E_{25}$ | 00:34:32 to 00:37:28 | 'What do you think – what factors will be the most influential [on the future development of MEDIN], do you think they'll be legal – changes in legal policy, *um* uses – use of like Creative Commons licences making things more open – social issues or technological issues?' | Yes |
| $E_{26}$ | 00:37:28 to 00:39:12 | '*Er* and do you think it good – it's good that MEDIN acts as a portal instead of just a centralised system where you [DASSH] sort of give all the data to them, and you just – you can have your own data centres tailored to what you want?' | Yes |
| $E_{27}$ | 00:39:12 to 00:40:53 | '*Um* what do you think other disciplines can learn from MEDIN and how – how marine data is being *um* being made re-usable? For example, legal data or chemistry data?' | Yes |
| $E_{28}$ | 00:40:53 to 00:41:53 | 'So do you think the government influence has been quite good for marine data whereas, say, other academic disciplines may not have that push?' | Yes |
| $E_{29}$ | 00:41:53 to 00:43:55 | '*Um* a lot of marine data before [the Web], about species [and] habitats, would – would it have been lost in the paper form? Or do you [DASSH] have a lot of back-dated records?' | Yes |
| $E_{30}$ | 00:43:55 to 00:45:08 | **9) MEDIN) [Ch3-S3.4.5-T2]** | No |
| $E_{31}$ | 00:45:08 to 00:46:04 | **10) [Ch3-S3.4.5-T2]** | No |
| **Question 3) not asked as this was added after the MEDIN case study interviews took place. See Chapter 3: Methodology, section 3.4.5 for further information.** | | | |
| *Interview completed at 00:46:04* | | | |

### B.1.3 Mr N. record of interview questions

| **Mr N.** Record of Interview Questions | | | |
|---|---|---|---|
| **Interview Section:** | **Question Duration:** | **Question:** | **Directly Referenced in thesis?** |
| $N_1$ | 00:00:00 to 00:01:26 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $N_2$ | 00:01:26 to 00:03:00 | **2) i) [Ch3-S3.4.5-T2]** | Yes |
| $N_3$ | 00:03:00 to 00:11:24 | **4) [Ch3-S3.4.5-T2]** | Yes |
| $N_4$ | 00:11:24 to 00:12:31 | '*Um* so many of the old maps and charts you may have from the beginning of the Hydrographic Office, are they available online?' | Yes |
| $N_5$ | 00:12:31 to 00:12:54 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $N_6$ | 00:12:54 to 00:15:10 | **6) [Ch3-S3.4.5-T2]** | Yes |
| $N_7$ | 00:15:10 to 00:15:38 | '*Um* did you [the Hydrographic Office] have your own standards before you joined MEDIN? How has that changed?' | No |
| $N_8$ | 00:15:38 to 00:16:33 | '*Um* so have *um* MEDIN taken from that [existing standards]?' | No |
| $N_9$ | 00:16:33 to 00:18:56 | **7) [Ch3-S3.4.5-T2]** | Yes |
| $N_{10}$ | 00:18:56 to 00:22:01 | 'So is this [bathymetric] data freely accessible to everyone online? […] It's not just giving the metadata to say it exists?' | Yes |
| $N_{11}$ | 00:22:01 to 00:23:31 | 'Do you think the roles of these portals [MEDIN, European INSPIRE, data.gov.uk] are really useful? Or do you think with Google [and other search engine providers] now, you can just – more people will just google the data instead of going to a portal?' | No |
| $N_{12}$ | 00:23:31 to 00:25:38 | **8) [Ch3-S3.4.5-T2]** | No |
| $N_{13}$ | 00:25:38 to 00:26:28 | '*Um* do you think many people are aware of MEDIN at the moment in the marine community?' | Yes |
| $N_{14}$ | 00:26:28 to 00:27:09 | **9) MEDIN) [Ch3-S3.4.5-T2]** | Yes |
| $N_{15}$ | 00:27:09 to 00:29:11 | 'And what about licensing considerations? People who give you [Hydrographic Office] their data, obviously you have a licence for how you handle that, does that have any impact on what you can make available?' | No |
| $N_{16}$ | 00:29:11 to 00:30:16 | 'And so, out of these raw datasets you [Hydrographic Office] make your own products? […] Yes. And then you'll have licences *um* for their use as well?' | No |
| $N_{17}$ | 00:30:16 to 00:30:35 | '*Um* so is your licensing inconsistent? Do you tend to use the open government licence, because obviously there's [sic] different types – Creative Commons …? | No |
| $N_{18}$ | 00:30:35 to 00:34:41 | '*Um* when did the Hydrographic Office first think about opening up dataset[s] on the Web, or even before the Web? Have there been any … examples?' | No |
| $N_{19}$ | 00:34:41 to 00:34:54 | '*Um* would you [Hydrographic Office] say 'cause you're sort of a data quality assurer – and making sure you've got this data that's good quality, and you can it into your products?' | No |
| $N_{20}$ | 00:34:54 to | '*Um* do you think INSPIRE's placed a burden on the | Yes |

| | | | |
|---|---|---|---|
| | 00:36:25 | Hydrographic Office and [other] organisations? Or do you think it's a good thing – getting people to open up sets – datasets?' | |
| $N_{21}$ | 00:36:25 to 00:37:08 | 'Do you think that's where MEDIN can maybe step in as a role [to support data reformatting] and …?' | No |
| $N_{22}$ | 00:37:08 to 00:39:22 | '*Um* because some people might say: 'that oh, this is public data, why do we need to pay for charts and maps?' But the thing, I suppose, [some] people forget is that people are being paid to make sure that it's quality, and that data products are made, and they're archived for a long time?' | No |
| $N_{23}$ | 00:39:22 to 00:39:30 | **10) [Ch3-S3.4.5-T2]** | No |
| **Question 3) not asked as this was added after the MEDIN case study interviews took place. See Chapter 3: Methodology, section 3.4.5 for further information. Due to human error, Question 2) ii) was not asked directly.** | | | |
| *Interview completed at 00:39:30* | | | |

### B.1.4 Dr S. record of interview questions

| **Dr S.** Record of Interview Questions | | | |
|---|---|---|---|
| **Interview Section:** | **Question Duration:** | **Question:** | **Directly Referenced in thesis?** |
| $S_1$ | 00:00:00 to 00:01:20 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $S_2$ | 00:01:20 to 00:02:20 | **2) i) [Ch3-S3.4.5-T2]** | Yes |
| $S_3$ | 00:02:20 to 00:04:31 | **2) ii) [Ch3-S3.4.5-T2]** | Yes |
| $S_4$ | 00:04.31 to 00:05.13 | '*Um* do you think MEDIN is [an] essential portal now for – the marine sciences? Or it's becoming [an] essential portal?' | Yes |
| $S_5$ | 00:05.13 to 00:05.40 | 'Are a lot of your colleagues [in the marine sciences] like that as well [they are more likely to use a search engine to find data on the Web rather than MEDIN directly]?' | Yes |
| $S_6$ | 00:05.40 to 00:07.15 | '*Um* with people in marine science, do you think there's a culture of sharing data within academia?' | Yes |
| $S_7$ | 00:07.15 to 00:07.57 | 'What about [academics sharing data] for commercial use?' | Yes |
| $S_8$ | 00:07.57 to 00:09:13 | '*Um* so from an academic's point of view, do you think archiving at a data centre is important? The focus on a data centre rather than the portal *um* – because obviously there's *um* data centres that feed into MEDIN, they pick up – and they're [MEDIN] just acting as the portal and trying to get more data centres involved?' | Yes |
| $S_9$ | 00:09:13 to 00:10:33 | 'Has a lot of *um* marine data been lost in the past, before the Web and things?' | Yes |
| $S_{10}$ | 00:10:33 to 00:11:54 | **4) [Ch3-S3.4.5-T2]** | Yes |
| $S_{11}$ | 00:11:54 to 00:12:21 | '*Um* do you think NERC has a big impact on – on oceanography, because they mandate that, you know, you [academics] have to deposit your data?' | Yes |
| $S_{12}$ | 00:12:21 to 00:13:21 | '*Um* do you think sometimes academics may feel overburdened if they have to fill in so many fields for provenance metadata, perhaps and – or they just feel it's just part of the culture so you do just go and deposit it?' | Yes |
| $S_{13}$ | 00:13:21 to 00:13:32 | **5) [Ch3-S3.4.5-T2]** | No |
| $S_{14}$ | 00:13:32 to 00:14:11 | 'Do – what do you *er* the legal issues do you think impact on your own research – your own data – do you feel …?' | Yes |
| $S_{15}$ | 00:14:11 to 00:15:08 | '*Um* thinking about copyright *um* and moral rights, authorial assertion is really important it's – that links into acknowledgement, so I suppose that's one area where …?' | No |
| $S_{16}$ | 00:15:08 to 00:15:17 | 'In physical oceanography have you never come across any – for example you can't re-use a map or anything because there are IPR restrictions?' | No |
| $S_{17}$ | 00:15:17 to 00:16:07 | **6)** [[…] **Dr S.** asks: 'When you say provenance metadata, do you mean – do you mean information about the dataset or information about who's collected it?' **LG** responds: '*Um*, information about dataset that includes who's collected it, and legal restrictions of | Yes |

| | | | |
|---|---|---|---|
| | | use and things.'] **[Ch3-S3.4.5-T2]** | |
| $S_{18}$ | 00:16:07 to 00:16:17 | 'Have you ever known any cases where *um* people have maybe not acknowledged people and that's led to some friction or anything? – No.' | No |
| $S_{19}$ | 00:16:17 to 00:19:21 | **7) [Ch3-S3.4.5-T2]** | Yes |
| $S_{20}$ | 00:19:21 to 00:20:34 | 'So what do you think sparked that [culture of sharing]? Do you think it's been technological innovation, where you can get more data automatically maybe and …?' | Yes |
| $S_{21}$ | 00:20:34 to 00:21:02 | '*Um* what about negative results – are those published as well? 'Cause in some scientific disciplines there's sort of a culture against not publishing those … […] Or, where things have gone wrong maybe?' | No |
| $S_{22}$ | 00:21:02 to 00:22:18 | 'Is that valuable to the community to [publish those unexpected results] …?' | No |
| $S_{23}$ | 00:22:18 to 00:23:58 | '*Um* if you have data in the past, maybe you collect a lot of data, but you haven't got time to write all the papers about every single [data]set so maybe it stays in your cabinet. But now even though it's unpublished you can still send it all maybe to a d – a data archive centre and they can go – it can go through quality review and things …?' | Yes |
| $S_{24}$ | 00:23:58 to 00:26:39 | **8) [Ch3-S3.4.5-T2]** | Yes |
| $S_{25}$ | 00:26:39 to 00:27:46 | **9) MEDIN) [Ch3-S3.4.5-T2]** | Yes |
| $S_{26}$ | 00:27:46 to 00:29:34 | '*Um* do you think government policy, sort of you've got the INSPIRE directive and things like that, have had a major impact on how marine data's re-used, collected …? […] It doesn't really impact on your research in physical oceanography' | Yes |
| $S_{27}$ | 00:29:34 to 00:30:15 | 'And *um* do you think that data re-usability in [physical] oceanography is already effective and appropriate, do you think it could be changed in anyway?' | No |
| $S_{28}$ | 00:30:15 to 00:31:03 | 'And *um* do you think that other disciplines could learn from [physical] oceanography about how they've handled their data and made it more re-usable? What sort of principles could they draw on?' | No |
| $S_{29}$ | 00:31:03 to 00:32:11 | '*Um* do you think MEDIN is well publicised within the [physical] oceanography community? […] And do you think if it was then, maybe, more people would go and use it or do you think …?' | Yes |
| $S_{30}$ | 00:32:11 to 00:32:18 | **10) [Ch3-S3.4.5-T2]** | No |
| **Question 3) not asked as this was added after the MEDIN case study interviews took place. See Chapter 3: Methodology, section 3.4.5 for further information.** | | | |
| *Interview completed at 00:32:18* | | | |

## B.1.5 Mrs T. record of interview questions

| **Mrs T.** Record of Interview Questions | | | |
|---|---|---|---|
| **Interview Section:** | **Question Duration:** | **Question:** | **Directly Referenced in thesis?** |
| $T_1$ | 00.00.00 to 00.01:22 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $T_2$ | 00.01:22 to 00.02:17 | '*Um* so do you have direct involvement with MEDIN?' | Yes |
| $T_3$ | 00.02:17 to 00.03:00 | **2) i) [Ch3-S3.4.5-T2]** | Yes |
| $T_4$ | 00:03:00 to 00:03:50 | 'So would you say it's better than just using Google [or another search engine], for example, just to search randomly for a dataset? It's better just to go to a focused portal?' | Yes |
| $T_5$ | 00:03:50 to 00:04:13 | **4) [Ch3-S3.4.5-T2]** | No |
| $T_6$ | 00:04:13 to 00:08:00 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $T_7$ | 00:08:00 to 00:08:54 | '*Um* so do the original data collectors, do they *um* maintain their copyright or do they assign it to you as the Hydrographic Office?' | Yes |
| $T_8$ | 00:08:54 to 00:10:00 | 'Does this make – has the *sui generis* database right made it quite confusing for people who maybe don't have an understanding of IP? When they're collecting data they don't know whether they should have a licence over it or a database right, they're not sure if copyright subsists?' | Yes |
| $T_9$ | 00:10:00 to 00:11:14 | '*Um* what about assertion of moral rights? Do you have that a lot? Is that very important?' | No |
| $T_{10}$ | 00:11:14 to 00:12:50 | 'Not even in the provenance metadata, you wouldn't have, I don't know: 'this person collected this data' …?' | No |
| $T_{11}$ | 00:12:50 to 00:13:19 | 'I was just thinking maybe that's [authorial assertion] different for academics, because obviously they thrive on publishing, and they want their name to be associated with data …?' | No |
| $T_{12}$ | 00:13:19 to 00:14:54 | '*Um* before the Open Government Licence, what was your [Hydrographic Office] copyright – what licensing did you have?' | Yes |
| $T_{13}$ | 00:14:54 to 00:15:24 | '*Um* what was the impact of Creative Commons? Or were you [Hydrographic Office] already – did you have these licences in existence before Creative Commons?' | Yes |
| $T_{14}$ | 00:15:24 to 00:17:56 | '*Um* do you [Hydrographic Office] have big problems with any third party copyright, or anything like that, when you're making your products or …?' | No |
| $T_{15}$ | 00:17:56 to 00:19:19 | '*Um* what about maps and charts that have been created here that are outside of copyright, from the 1700s – can – and they haven't been made available? Is there any provision for them to be made available online?' | No |
| $T_{16}$ | 00:19:19 to 00:22:23 | **6) [Ch3-S3.4.5-T2]** | Yes |
| $T_{17}$ | 00:22:23 to 00:23:39 | 'Just think it's – you're signposting that this dataset exists on *um* MEDIN – it's also signposting, oh, these are actually the rights, these are the restrictions and what you can do with this data that actually make it very clear to the user what they can do and maybe before it hasn't been so clear? You have to maybe | Yes |

| | | | |
|---|---|---|---|
| | | release – release the whole licence?' | |
| $T_{18}$ | 00:23:39 to 00:25:06 | '*Um* 'cause what I was hearing […] is that there's [sic] a lot of problems with people using different licensing arrangements at the moment [in the marine community]? Some people are just using bespoke there's – there's a lot of bespoke licensing going on, there's the open government licence? Would it be better if MEDIN had one licence they used for all their data? If they sort of mandated that everyone use one licence?' | Yes |
| $T_{19}$ | 00:25:06 to 00:26:51 | '*Um* do you think the marine community is one that's keen to share its data with the general public as widely available as possible?' | No |
| $T_{20}$ | 00:26:51 to 00:28:37 | **7) [Ch3-S3.4.5-T2]** | Yes |
| $T_{21}$ | 00:28:37 to 00:30:22 | '*Um* what impact has the INSPIRE Directive had on making datasets available [at the Hydrographic Office]?' | Yes |
| $T_{22}$ | 00:30:22 to 00:30:58 | 'Has there been a long history here [Hydrographic Office] of making data available, or has it just been a very recent thing?' | No |
| $T_{23}$ | 00:30:58 to 00:32:51 | **8) [Ch3-S3.4.5-T2]** | Yes |
| $T_{24}$ | 00:32:51 to 00:34:02 | '*Um* do you think people would be more likely to come to your [Hydrographic Office] website, or if it was on data.gov, there, or MEDIN as portal, or just Google [or another search engine], to get the dataset they wanted?' | Yes |
| $T_{25}$ | 00:34:02 to 00:34:58 | '*Um* when it comes to raw data at the Hydrographic Office, your – you obviously have crown copyright, but some of it – are you just sort of a data guardian, and use it in your products? You've licensed it from the third parties for [re-]use?' | No |
| $T_{26}$ | 00:34:58 to 00:35:13 | **9) MEDIN) [Ch3-S3.4.5-T2]** | No |
| $T_{27}$ | 00:35:13 to 00:36:43 | **10) [Ch3-S3.4.5-T2]** | Yes |
| $T_{28}$ | 00:36:43 to 00:37:07 | 'Is that because MEDIN is new [as there is still a lot of data that MEDIN is not collecting] or do you think they [MEDIN] need to publicise themselves more?' | Yes |
| $T_{29}$ | 00:37:07 to 00:37:13 | '*Um* is there anything else?' | No |
| **Question 3) not asked as this was added after the MEDIN case study interviews took place. See Chapter 3: Methodology, section 3.4.5 for further information. Due to human error, Question 2) ii) was not asked directly.** | | | |
| *Interview completed at 00:37:13* | | | |

### B.1.6 Mr W. record of interview questions

| **Mr W.** Record of Interview Questions | | | |
|---|---|---|---|
| **Interview Section:** | **Question Duration:** | **Question:** | **Directly Referenced in thesis?** |
| $W_1$ | 00:00:00 to 00:00:47 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $W_2$ | 00.00.47 to 00.01.20 | **2) i) [Ch3-S3.4.5-T2]** | Yes |
| $W_3$ | 00.01.20 to 00.04.13 | **2) ii) [Ch3-S3.4.5-T2]** | Yes |
| $W_4$ | 00.04.13 to 00:04:31 | 'So is this a bit like linked data – sort of Semantic Web?' | No |
| $W_5$ | 00.04.31 to 00.05.13 | '*Um* what are the plans for the future to improve on this [MEDIN and BODC's limited metadata search capabilities]?' | Yes |
| $W_6$ | 00.05.13 to 00.06.39 | **4) [Ch3-S3.4.5-T2]** | Yes |
| $W_7$ | 00.06.39 to 00.07.48 | 'I was thinking in different disciplines sometimes when *um* datasets may have been published – the publisher may say: 'oh well, we just want to keep that dataset and people have to pay us to view it [through subscription charges].' But if, sort of, an academic has published a dataset then – and it appears on MEDIN that's not really an issue with marine data?' | Yes |
| $W_8$ | 00.07.48 to 00.09.27 | 'Do you think it [depositing data at data archive centres] may help *um* journal publishing? Because, I've been reading that sometimes you never – you don't see the raw data when it's published, because people want to read stats and things, which sort of hide [raw data] everything behind. But with MEDIN, if you can have a citation back to the data, and have the provenance – then that's going to help?' | Yes |
| $W_9$ | 00.09.27 to 00.11.00 | 'Also in the past, has a lot of marine data been lost, because there just wasn't the Web or these kind of [thematic data] archive centres, in the – in the paper form?' | Yes |
| $W_{10}$ | 00.11.00 to 00.12.10 | '*Er* with marine data do you find there's an overload of data as we hear on the Web there's 'a big overload of data' – or do you actually see – find that the marine data is – it's scarce and you want to encourage more data to be produced?' | No |
| $W_{11}$ | 00.12.10 to 00.14.45 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $W_{12}$ | 00.14.45 to 00.15.19 | 'So self-regulation is better than pursuing legal avenues really? If someone's broken your licence agreement.' | Yes |
| $W_{13}$ | 00.15.19 to 00:18.11 | '*Um* what's the impact of copyright, would you say, on data re-usability – sort of to do with licensing and things? Does it help/hinder what you [MEDIN/BODC] can do with your datasets?' | Yes |
| $W_{14}$ | 00:18.11to 00:19.05 | 'What do you think could've been done with that project to encourage people, well, to allow people to use it and copy it? Do you think that's a legal issue that needs to be addressed or is it more of a social problem – where people just want to hold onto their – their data or information and don't want to hand it over?' | No |
| $W_{15}$ | 00:19.05to 00:20.22 | '*Um* do you think that licensing is still useful though for data?' | Yes |

| | | | |
|---|---|---|---|
| $W_{16}$ | 00:20.22 to 00:21.24 | 'Um 'cause following on from that, do you think a lot of academic data – research data doesn't have a commercial value then? Is a lot of it just for research that's publically funded, so it wouldn't be exploited commercially?' | No |
| $W_{17}$ | 00:21.24 to 00.22.44 | **6) [Ch3-S3.4.5-T2]** | Yes |
| $W_{18}$ | 00:22.44 to 00:24.14 | 'Um so do you think the legal boundaries of use is [sic] an essential part of provenance metadata? Should be an essential feature?' | Yes |
| $W_{19}$ | 00:24.14 to 00:24.52 | 'Um how much of the – the data on MEDIN is embargoed, or is embargoed indefinitely or there's restrictions? And how much is open to anyone – any member of the public at the moment?' | Yes |
| $W_{20}$ | 00:24.52 to 00:25.04 | 'And are their metadata still exposed even if [these data held by the thematic data archive centres] they're completely restricted?' | Yes |
| $W_{21}$ | 00:25.04 to 00:26.32 | 'Um so do you think that's one good thing [signposting data] from the Web, is that *um* because now we have [digital] archives. Before, perhaps, someone had some data locked in a cupboard, and you don't [as a research user] have anything even the metadata to flag that it exists. But now, at least, even if it – it's restricted you know what – you have the metadata. You know that it exists. So, maybe you can contact the person [rights holder] to sort of see if they'll – they'll allow you to use it?' | No |
| $W_{22}$ | 00:26.32 to 00.28.56 | **7) [Ch3-S3.4.5-T2]** | Yes |
| $W_{23}$ | 00:28.56 to 00:32.35 | **8) [Ch3-S3.4.5-T2]** | Yes |
| $W_{24}$ | 00:32.35 to 00:33.44 | 'Um so from that what – what issues do you think impact on the running of MEDIN the most? Are they the social issues, sort of academics or researchers' awareness of *um* depositing their data; or, is it socio-legal, the awareness about the legal rights; or, is it more about the technical issues that we've just discussed?' | Yes |
| $W_{25}$ | 00:33.44 to 00:35.32 | 'Have you found that the user base has grown? Or can you track sort of *erm* researchers from this company or this university using MEDIN?' | Yes |
| $W_{26}$ | 00:35.32 to 00.36.16 | **9) MEDIN) [Ch3-S3.4.5-T2]** | No |
| $W_{27}$ | 00.36.16 to 00:36.42 | 'Or is that not a question [9) MEDIN)] maybe for MEDIN, but who – the funding bodies and things?' | Yes |
| $W_{28}$ | 00:36.42 to 00:39.30 | 'Um what else I wanted to ask is: what do you think other areas – academic areas or disciplines can learn from MEDIN and how the portal operates? And perhaps the policy of – the NERC policy as well, you know they mandate that data should be put with MEDIN? *Um* could, I don't know, chemists or other people – lawyers, use sort of similar principles they can take from MEDIN and build their own systems?' | No |
| $W_{29}$ | 00:39.30 to 00:40.23 | 'Um and what are the key things, would you say, that make data re-usable?' | No |
| $W_{30}$ | 00:40.23 to 00:42.08 | 'Um with the metadata, some may say it could be burdensome if you've got lots and lots of metadata and hard to trawl through, sort of, how do you, sort of, work with that at MEDIN?' | No |
| $W_{31}$ | 00:42.08 to | 'Um in some sciences *er* scientific disciplines you | No |

| | | | |
|---|---|---|---|
| | 00:42.53 | might get negative results, which in the past perhaps people haven't been keen about publishing. But *um*, at MEDIN do you get sort of results from, say I don't know, experiments or data collection where it's not quite what they wanted?' | |
| $W_{32}$ | 00:42.53 to 00:44.18 | '*Um* and I know it's quite a generalisation, but *um* do you think in marine sciences the culture is to share? Or do you think in the past it's changed, because we've had the Web, and people now think: 'yeah, we can put it in MEDIN and, you know, share the data or even disclose that it exists?' | Yes |
| $W_{33}$ | 00:44.18 to 00:44.49 | **10) [Ch3-S3.4.5-T2]** | No |
| $W_{34}$ | 00:44.49 to 00:45.41 | 'Actually that's – I'm sorry, that's one point *um* with *er* embargoing and things, sometimes they say they're indefinitely embargoed and there's no plan, even say in a hundred years' time, to have a policy where you can open things up. I was just wondering if people at these data centres, they're thinking that far ahead for future generations, if they can say: 'well this might not be restricted forever, but it might be another generation – if it's like highly sensitive information or something?' | No |
| $W_{35}$ | 00:45.41 to 00:46:12 | 'Is that more of an issue [indefinite embargoes] with people from commercial bodies?' | No |
| $W_{36}$ | 00:46:12 to 00:46.19 | '*Er* is there anything else?' | No |
| **Question 3) not asked as this was added after the MEDIN case study interviews took place. See Chapter 3: Methodology, section 3.4.5 for further information.** | | | |
| *Interview completed at 00:46:19* | | | |

# B.2 Chapter 4: eCrystals and LabTrove case study

## B.2.1 Dr A. record of interview questions

| **Dr A.** Record of Interview Questions | | | |
|---|---|---|---|
| **Interview Section:** | **Question Duration:** | **Question:** | **Directly Referenced in thesis?** |
| $A_1$ | 00:00:00 to 00:01:22 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $A_2$ | 00:01:22 to 00:02:50 | **2) i) [Ch3-S3.4.5-T2]** | Yes |
| $A_3$ | 00:02:50 to 00:04:50 | **2) ii) [Ch3-S3.4.5-T2]** | Yes |
| $A_4$ | 00:04:50 to 00:06:20 | **3) [Ch3-S3.4.5-T2]** | Yes |
| $A_5$ | 00:06:20 to 00:06:36 | '*Um* do you have any like *um* data warnings or things like that, that you use? [A] data flag system or anything?' | No |
| $A_6$ | 00:06:36 to 00:08:00 | '*Um* how do you think a thing [sic] like eCrystals – Electronic Lab – Lab Notebook can support traditional *um* publishing processes?' | Yes |
| $A_7$ | 00:08:00 to 00:10:48 | **4) [Ch3-S3.4.5-T2]** | Yes |
| $A_8$ | 00:10:48 to 00:11:18 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $A_9$ | 00:11:18 to 00:11:36 | '*Um* what about this – *um* the main legal issues with the Electronic Lab Notebook?' | Yes |
| $A_{10}$ | 00:11:36 to 00:11:48 | '*Um* what about the licensing – the data licensing on eCrystals?' | Yes |
| $A_{11}$ | 00:11:48 to 00:13:38 | **6)** [**Dr A.** asks: 'Well, you know, I'm not quite sure what you mean by provenance metadata?' **LG** responds: 'So, anything – all the information about the origins, the chain of custody, *um* the quality assurance [of] data – of a particular datum …'] **[Ch3-S3.4.5-T2]** | Yes |
| $A_{12}$ | 00:13:38 to 00:14:18 | '*Um*, so is it based on trust? […] There's no independent quality check?' | Yes |
| $A_{13}$ | 00:14:18 to 00:15:18 | '*Um* what about the embargo system?' | Yes |
| $A_{14}$ | 00:15:18 to 00:15:33 | '*Um* are there any plans to rectify that [the limited search function on eCrystals]?' | Yes |
| $A_{15}$ | 00:15:33 to 00:16:46 | **7) [Ch3-S3.4.5-T2]** | Yes |
| $A_{16}$ | 00:16:46 to 00:17:16 | '*Um* do you track the number of users that use eCrystals?' | Yes |
| $A_{17}$ | 00:17:16 to 00:17:40 | 'So really the plans to expand in the future – is it just 'cause it's early days for eCrystals?' | Yes |
| $A_{18}$ | 00:17:40 to 00:18:21 | 'So funding is really a big part of data management?' | Yes |
| $A_{19}$ | 00:18:21 to 00:20:38 | **8) [Ch3-S3.4.5-T2]** | Yes |
| $A_{20}$ | 00:20:38 to 00:21:10 | 'Do you think there is a bigger audience of research users who would like to use something like eCrystals – if it expanded?' | Yes |
| $A_{21}$ | 00:21:10 to 00:22:35 | '*Um* what do you think about the use of linked data?' [**Dr A.** asks: 'So, *um* I don't really know what you mean by that?' **LG** responds: 'So, using sort of a linked data system where – I was thinking – when you | No |

| | | | |
|---|---|---|---|
| | | have data structures you can have *um* the URIs, so when it does come to publications you've got a definitive version that you can link back to and you sort of …'] | |
| $A_{22}$ | 00:22:35 to 00:24:06 | 'Do you think that's a problem for scientific disciplines at the moment, sort of – there's a lot of multi-authorship that seems to keep on rising? And you [potentially] have sort of ghost authors – ghost authors and things like that sometimes?' | Yes |
| $A_{23}$ | 00:24:06 to 00:24:38. | **9) i) [Ch3-S3.4.5-T2]** | Yes |
| $A_{24}$ | 00:24:38 to 00:25:32 | **9) ii) [Ch3-S3.4.5-T2]** | Yes |
| $A_{25}$ | 00:25:32 to 00:26:27 | '*Um* how do you think chemistry research has changed in recent years – if it has?' | Yes |
| $A_{26}$ | 00:26:27 to 00:27:51 | **9) iii) [Ch3-S3.4.5-T2]** | Yes |
| $A_{27}$ | 00:27:51 to 00:29:05 | 'How can the use of the Electronic Lab Notebook influence this [quality assurance] as well?' | No |
| $A_{28}$ | 00:29:05 to 00:29:32 | '*Um* what do you think about the access – the open access movement on the Web? Do you think that's had an impact on chemistry?' | No |
| $A_{29}$ | 00:29:32 to 00:30:51 | 'Is it more of a top-down push from funding bodies telling academics what they should be doing with their data [i.e. mandating open access] rather than academics sort of taking it upon themselves [to make their data open]?' | No |
| $A_{30}$ | 00:30:51 to 00:31:13 | '*Um* when the publishers in chemistry – do they actually review the – the raw data themselves – you generally or …?' | Yes |
| $A_{31}$ | 00:31:13 to 00:32:23 | '*Um* is there a problem with data duplication in chemistry? Because, if some – sometimes things are closed people might go and do the same experiment?' | Yes |
| $A_{32}$ | 00:32:23 to 00:32:33 | '*Um* what about negative results?' | No |
| $A_{33}$ | 00:32:33 to 00:33:07. | 'But do you think it would be beneficial if they [negative results] were published – so people knew what goes wrong?' | No |
| $A_{34}$ | 00:33:07 to 00:34:14 | **10) [Ch3-S3.4.5-T2]** | Yes |
| *Interview completed at 00:34:14* | | | |

## B.2.2 Mr C. record of interview questions

| **Mr C.** Record of Interview Questions | | | |
|---|---|---|---|
| Interview Section: | Question Duration: | Question: | Directly Referenced in thesis? |
| $C_1$ | 00:00:00 to 00:03:06 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $C_2$ | 00:03:06 to 00:05:03 | **2) i) [Ch3-S3.4.5-T2]** | Yes |
| $C_3$ | 00:05:03 to 00:08:14 | '*Um* do you think all data at the University should be put into ePrints? Or do you think having *er* separate discipline portals of data – their own repositories – is a good idea?' | Yes |
| $C_4$ | 00:08:14 to 00:11:22 | **2) ii) [Ch3-S3.4.5-T2]** | Yes |
| $C_5$ | 00:11:22 to 00:11:53 | 'Do you think open software methods are better than say more bespoke systems that are more expensive, or …?' | No |
| $C_6$ | 00:11:53 to 00:14:43 | **3) [Ch3-S3.4.5-T2]** | Yes |
| $C_7$ | 00:14:43 to 00:17:20 | **4) [Ch3-S3.4.5-T2]** | Yes |
| $C_8$ | 00:17:20 to 00:18:50 | '*Um* what do you think – how do think it [academic research data re-usage models] helps the publishers' quality assurance processes? If you've got say another copy of that data that's made open on the Web that other people go and scrutinise it? Whereas, there've been cases where *um* fraudulent articles have been published, because maybe the data haven't been quality assessed by the publishing process as much.' | Yes |
| $C_9$ | 00:18:50 to 00:19.31 | 'Do you think *um* some – some academics creating research data are concerned [about] making their data open, because they're worried that it might be exposed that it's not quite as high quality or there's some errors in it?' | Yes |
| $C_{10}$ | 00:19:31 to 00:22:11 | **5) [Ch3-S3.4.5-T2]** | No |
| $C_{11}$ | 00:22:11 to 00:24:30. | '*Um* how do you think data licensing has an impact on the re-use of data? Sort of Creative Commons and open government licensing – do you think they've, sort of – they've sort of meshed together well?' | Yes |
| $C_{12}$ | 00:24:30 to 00:25:29 | **6) [Ch3-S3.4.5-T2]** | Yes |
| $C_{13}$ | 00:25:29 to 00:27:51 | **7) [Ch3-S3.4.5-T2]** | Yes |
| $C_{14}$ | 00:27:51 to 00:30:20 | '*Um* following on with that what do you think about the Finch Report? Sort of this top-down push telling research councils – mandating that all data be made accessible where it can be?' | Yes |
| $C_{15}$ | 00:30:20 to 00:31:27 | '*Um* also how do you think sort of open data – open research data – is helping citizen science and sort of the democratisation of knowledge?' | No |

| | | | |
|---|---|---|---|
| $C_{16}$ | 00:31:27 to 00:32:57 | '*Um* do you think it's helping the humanities in any way? Sort of open data, because you hear a lot about the push towards the sciences, but not so much when it comes to the arts and the humanities?' | No |
| $C_{17}$ | 00:32:57 to 00:33:47 | '*Um* what do you think sets academic research data apart from other types of data on the Web?' | No |
| $C_{18}$ | 00:33:47 to 00:35:36 | **8) [Ch3-S3.4.5-T2]** | Yes |
| $C_{19}$ | 00:35:36 to 00:36:12 | '*Um* what role do you think linked data will play in this?' | Yes |
| $C_{20}$ | 00:36:12 to 00:36:51 | 'Do you think *um* this [linked data] *um* it will change quality assurance processes?' | No |
| $C_{21}$ | 00:36:51 to 00:37:42 | 'Do you see a lot of – a lot more data being made open – especially things that [came] before the advent of the Web are not born digital?' | No |
| $C_{22}$ | 00:37:42 to 00:38:50 | **9) i) [Ch3-S3.4.5-T2]** | Yes |
| $C_{23}$ | 00:38:50 to 00:39:52 | **9) ii) [Ch3-S3.4.5-T2]** | Yes |
| $C_{24}$ | 00:39:52 to 00:41:20 | '*Um* what about Web delivery of data? *Um* do you think it's a misconception that people think they can just go and access all the data digitally, instantaneously? *Er*, people still need to send sort of hard drives and things through the post, or ring up and ask that they can go and get a paper copy?' | Yes |
| $C_{25}$ | 00:41:20 to 00:42:03 | **9) iii) [Ch3-S3.4.5-T2]** | Yes |
| $C_{26}$ | 00:42:03 to 00:42:17 | **10) [Ch3-S3.4.5-T2]** | No |
| *Interview completed at 00:42:17* | | | |

### B.2.3 Dr G. record of interview questions

| **Dr G.** Record of Interview Questions | | | |
|---|---|---|---|
| Interview Section: | Question Duration: | Question: | Directly Referenced in thesis? |
| $G_1$ | 00.00.00 to 00.01.59 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $G_2$ | 00.01.59 to 00.05.09 | **2) i) [Ch3-S3.4.5-T2]** | Yes |
| $G_3$ | 00.05.09 to 00.07.43 | '*Um* do you record your [eCrystal's] research users? Do you know who they are – who's accessing [data]?' | Yes |
| $G_4$ | 00.07.43 to 00.12.09 | **2) ii) [Ch3-S3.4.5-T2]** | Yes |
| $G_5$ | 00.12.09 to 00.14.39 | **3) [Ch3-S3.4.5-T2]** | Yes |
| $G_6$ | 00.14.39 to 00.16.49 | '*Um* 'cause I was looking on some of the [eCrystals] records that the *um* – I think the points you give for – for different things, you know, on your *um* quality controls – the different colours – I was just wondering if you could expand more on them? 'Cause are they sort of like quality control flags?' | Yes |
| $G_7$ | 00.16.49 to 00.19.56 | **4) [Ch3-S3.4.5-T2]** | Yes |
| $G_8$ | 00.19.56 to 00.21.38 | 'What about the peer review process at journals – are they quite happy to, sort of, say the repository can do that rather than themselves or …?' | Yes |
| $G_9$ | 00.21.38 to 00:24:30 | '*Um*, how do they [eCrystals and the ELN] impact on funding bodies and academics as well?' | Yes |
| $G_{10}$ | 00:24:30 to 00:27:49 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $G_{11}$ | 00:27:49 to 00:32:14 | **6) [Ch3-S3.4.5-T2]** | Yes |
| $G_{12}$ | 00:32:14 to 00:33:04 | 'Also, with the vocabulary, do you sort of *um* have working groups for standardisation and metadata within chemistry or is it something made here?' | Yes |
| $G_{13}$ | 00:33:04 to 00:34:51 | **7) [Ch3-S3.4.5-T2]** | Yes |
| $G_{14}$ | 00:34:51 to 00:36:58 | 'Yep, so that's what the closed records are [on eCrystals] – things that are …?' | Yes |
| $G_{15}$ | 00:36:58 to 00:40:20 | **8) [Ch3-S3.4.5-T2]** | Yes |
| $G_{16}$ | 00:40:20 to 00:42:46 | **9) eCrystals and LabTrove [Ch3-S3.4.5-T2]** | Yes |
| $G_{17}$ | 00:42:46 to 00:43:35 | '*Um*, with chemistry, because there's a lot [of data] generated – I know we discussed this before about *erm* – how it's bypassing journal articles can be a good thing, because how long each journal article would take to write up. So you could sort of publish lots of data and not have to go through a formal process?' | Yes |
| $G_{18}$ | 00:43:35 to 00:43:57 | '*Um* and just *um* – do you see yourself at eCrystals as a – both a data author and data manager? So you're a custodian as well?' | Yes |
| $G_{19}$ | 00:43:57 to 00:47:49 | **10) [Ch3-S3.4.5-T2]** | Yes |
| *Interview completed at 00:47:49* | | | |

## B.2.4 Mr H. record of interview questions

| **Mr H.** Record of Interview Questions | | | |
|---|---|---|---|
| **Interview Section:** | **Question Duration:** | **Question:** | **Directly Referenced in thesis?** |
| $H_1$ | 00:00:00 to 00:00:50 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $H_2$ | 00:00:50 to 00:02:48 | **2) i) [Ch3-S3.4.5-T2]** | Yes |
| $H_3$ | 00:02:48 to 00:04:40 | '*Um* do you get a lot of people – do you get a lot of people from sciences, social sciences and humanities *um* coming to you for advice? Do you get many from the sciences or –? | Yes |
| $H_4$ | 00:04:40 to 00:06:31 | **2) ii) [Ch3-S3.4.5-T2]** | Yes |
| $H_5$ | 00:06:31 to 00:10:01 | **3) [Ch3-S3.4.5-T2]** | Yes |
| $H_6$ | 00:10:01 to 00:11:19 | **4) [Ch3-S3.4.5-T2]** | No |
| $H_7$ | 00: 11:19 to 00:14:52 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $H_8$ | 00: 14:52 to 00:18:09 | 'Do you think copyright law itself is confusing? – Because it's confusing whether you – copyright subsists in a dataset in the first place – or you have a database right. And then do you get some instances where some people maybe take out a data licence when they actually don't need one?' | Yes |
| $H_9$ | 00:18:09 to 00:19:37 | '*Um* … oh, *um* what do you think the role of 'moral rights' are? Do you think they're actually relevant when it comes to big *um* collaborative projects? – Where you have, sort of, ten people with one – creating one dataset – they may have mashed and merged other datasets together. Do you think that causes big problems?' | Yes |
| $H_{10}$ | 00:19:37 to 00:20:54 | 'So really contract law is king?' | No |
| $H_{11}$ | 00:20:54 to 00:23:48 | '*Um* what do you think the impact of open licensing – such as Creative Commons and the Open Government Licence – has had on data reuse?' | Yes |
| $H_{12}$ | 00:23:48 to 00:25:54 | '*Erm* do you think the problem, maybe, with *um* sort of copyright law and things – is that there's not much case law when it comes to sort of misuse of data, sort of moral rights and things? And, sort of, when you focus on law – it's more about non-contentious work about – it is contract drafting and it's that side of it. Rather than and – people are like: 'well there's been no case law – so I don't know if this actually would go to court even if I misused this data?' | Yes |
| $H_{13}$ | 00:25:54 to 00:28:13 | **6) [Ch3-S3.4.5-T2]** | Yes |
| $H_{14}$ | 00:28:13 to 00:32:11 | '*Um* if [Legal Services] you did have unlimited resources, money, staff and time – what tools would you create to sort of help facilitate data depositing?' | Yes |
| $H_{15}$ | 00:32:11 to 00:33:11 | '*Um* do you get asked – Legal Services get asked a lot to give seminars and talks around the University to inform – give general advice?' | No |
| $H_{16}$ | 00:33:11 to 00:34:22 | 'Would you say your [Legal Services'] expertise is, sort of, under-used in the University?' | Yes |

| | | | |
|---|---|---|---|
| $H_{17}$ | 00:34:22 to 00:37:24 | '*Um* do you think – I know this is very general – but do you think the law as it stands facilitates open data and data re-use? – Because, you could argue that copyright law wasn't really meant to sort of fuel this.' | Yes |
| $H_{18}$ | 00:37:24 to 00:41:09 | 'And I suppose because there's all these different categories in copyright law – there's no real legal definition of what is – what – what are 'data'. I mean you have *um* 'database', but it maybe – it may seem a bit out-dated now?' | No |
| $H_{19}$ | 00:41:09 to 00:43:53 | **7) [Ch3-S3.4.5-T2]** | Yes |
| $H_{20}$ | 00:43:53 to 00:45:26 | 'What do you think about liability for accuracy of datasets?' | Yes |
| $H_{21}$ | 00:45:26 to 00:45:40 | '*Um* do you think a disclaimer's required [about data quality] –?' | Yes |
| $H_{22}$ | 00:45:40 to 00:48:10 | **8) [Ch3-S3.4.5-T2]** | Yes |
| $H_{23}$ | 00:48:10 to 00:49:51 | 'Have you had any – well – or in a general sense, any problems with *um* – when you have a data owner, and maybe you can't actually trace them from years ago? But the University may hold a dataset?' | No |
| $H_{24}$ | 00:49:51 to 00:52:25 | '*Um* do the rights vest with the *um* academic who created the data or – so when they say – they move University, they still have rights, they can take things with them? Or, as it's made in the course of their employment, do[es] the University own the data?' | Yes |
| $H_{25}$ | 00:52:25 to 00:53:28 | **9) [Ch3-S3.4.5-T2]** | No |
| $H_{26}$ | 00:53:28 to 00:55:55 | '*Um* how has the Web impacted on your role [as a legal advisor at University of Southampton]?' | No |
| $H_{27}$ | 00:55:55 to 00:58:00 | **10) [Ch3-S3.4.5-T2]** | Yes |
| *Interview completed at 00:58:00* | | | |

## B.2.5 Miss J. record of interview questions

| **Miss J.** Record of Interview Questions | | | |
|---|---|---|---|
| **Interview Section:** | **Question Duration:** | **Question:** | **Directly Referenced in thesis?** |
| $J_1$ | 00:00:00 to 00:01:07 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $J_2$ | 00:01:07 to 00:03:25 | **2) i) [Ch3-S3.4.5-T2]** | Yes |
| $J_3$ | 00:03:25 to 00:04:32 | **2) ii) [Ch3-S3.4.5-T2]** | Yes |
| $J_4$ | 00:04:32 to 00:07:02 | **3) [Ch3-S3.4.5-T2]** | Yes |
| $J_5$ | 00:07:02 to 00:08:17 | '*Um* do you think a formal or informal system of quality assurance is best?' | Yes |
| $J_6$ | 00:08:17 to 00:11:52 | **4)** [[…] **Miss J.** asks: 'In terms of does it impact on publishers *er* this is *er*. I'm going to ask for a little bit of clarification here, because 'impact' kind of to me kind of has a negative connotation. Is that what –?' **LG** responds: 'No – a positive effect – can it support the publishing process as well – if you've got the definitive version –.'] **[Ch3-S3.4.5-T2]** | Yes |
| $J_7$ | 00:11:52 to 00:13:01 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $J_8$ | 00:13:01 to 00:13:44 | 'Do patents have an impact [on the re-use of crystallography data]?' | Yes |
| $J_9$ | 00:13:44 to 00:15:42 | '*Um* … What do you think about data licensing as well? [**Miss J.** asks: '… *Er* so in what context with respect to – are we talking about sort of Creative Commons licensing or–?' **LG** responds: 'Yeah. Any open type of licensing or –?'] | Yes |
| $J_{10}$ | 00:15:42 to 00:18:04 | '*Um* do you think a lot of academics in Chemistry use the Legal Services and use legal advice from available outlets within the University? – If they're unclear about whether, you know, they should use a data licence or whether copyright even subsists in what they're doing?' | Yes |
| $J_{11}$ | 00:18:04 to 00:19:50 | **6) [Ch3-S3.4.5-T2]** | Yes |
| $J_{12}$ | 00:19:50 to 00:21:36 | '*Um* what do you think makes data re-usable?' [**Miss J.** asks: Well … Do you mean what is inherent about data that makes it so that it can be re-used or what is it – is it about the storage technique that makes it re-usable?' **LG** responds: 'If you had to have sort of a list of criteria you think well – if to make data [re-usable] […] what do I have to do?' | Yes |
| $J_{13}$ | 00:21:36 to 00:22:35 | '*Um* in crystallography has there been a lot of data loss over the – the years?' | Yes |
| $J_{14}$ | 00:22:35 to 00:22:45 | 'Has there also been a lot of data duplication [in crystallography]?' | No |
| $J_{15}$ | 00:22:45 to 00:24:13 | 'And what about negative results [in crystallography]? [**Miss J.** asks: 'What do you mean by that?' **LG** responds: '*Um* do they get readily published or do they get put in repositories to show – to show that this process didn't work or …?' | No |
| $J_{16}$ | 00:24:13 to 00:24:49 | **7) [Ch3-S3.4.5-T2]** | No |
| $J_{17}$ | 00:24:49 to 00:26:43 | **8) [Ch3-S3.4.5-T2]** | Yes |
| $J_{18}$ | 00:26:43 to 00:27:32 | '*Um* can you see mergers with other repositories, such as the one in Cambridge?' | Yes |
| $J_{19}$ | 00:27:32 to 00:30:28 | '*Um* do you think there can be a problem with portal multiplicity where you have lots of different versions of one | Yes |

| | | | |
|---|---|---|---|
| | | dataset? *Um*, because they're harvested maybe from different places, or if people deposit them in different repositories – I'm not sure if that's a big problem for chemistry?' | |
| $J_{20}$ | 00:30:28 to 00:34:13 | **9) [Ch3-S3.4.5-T2]** | Yes |
| $J_{21}$ | 00:34:13 to 00:35:44 | '*Um* do you think repositories in chemistry are learning from other repositories in different disciplines, or is it very much coming from that [chemistry] community?' | No |
| $J_{22}$ | 00:35:44 to 00:36:13 | '*Um* what do you think about data policy at the University [of Southampton]? [Also] *um* trying to make ePrints [Soton] an outlet for data?' | No |
| $J_{23}$ | 00:36:13 to 00:37:18 | 'Do you think there could be more legal advice made available?' | No |
| $J_{24}$ | 00:37:18 to 00:39:50 | '*Um* do you think you need a top-down push for data re-usability for people to make these repositories, for funding bodies to mandate that you should make your data open and provide channels to do this? Or, do you think it's down to academics themselves who decide that they – they like the – what *er* the benefits of open data?' | Yes |
| $J_{25}$ | 00:39:50 to 00:40:51 | '*Um* do you also think that it's down to education of getting, I don't know, masters students or even undergrads to understand open science, open data – the benefits that come from open access?' | No |
| $J_{26}$ | 00:40:51 to 00:41:09 | 'What about legal modules? Did you get any experience of that?' | No |
| $J_{27}$ | 00:41:09 to 00:41:17 | **10) [Ch3-S3.4.5-T2]** | No |
| *Interview completed at 00:41:17* | | | |

### B.2.6 Ms R. record of interview questions

| **Ms R.** Record of Interview Questions | | | |
|---|---|---|---|
| **Interview Section:** | **Question Duration:** | **Question:** | **Directly Referenced in thesis?** |
| $R_1$ | 00:00:00 to 00:01:44 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $R_2$ | 00:01:44 to 00:08:30 | **2) i) [Ch3-S3.4.5-T2]** | Yes |
| $R_3$ | 00:08:30 to 00:17:42 | 'Do you think that [open data collection tools and their preservation] would help non-technical people maybe *um* [some] people in the humanities [without this expertise] sort of academics?' | Yes |
| $R_4$ | 00:17:42 to 00:22:20 | **2) ii) [Ch3-S3.4.5-T2]** | Yes |
| $R_5$ | 00:22:20 to 00:27:32 | '*Um* what do you think about amateurs versus professionals when it comes to data collection and analysis on the Web? – Sort of citizen science?' | Yes |
| $R_6$ | 00:27:32 to 00:28:53 | **3) i) ii) [Ch3-S3.4.5-T2]** | No |
| $R_7$ | 00:28:53 to 00:30:47 | 'Who deals with *um* the quality assurance procedures [within ePrints Soton]…?' | No |
| $R_8$ | 00:30:47 to 00:32:32 | **3) iii) [Ch3-S3.4.5-T2]** | Yes |
| $R_9$ | 00:32:32 to 00:35:03 | **4) [Ch3-S3.4.5-T2]** | Yes |
| $R_{10}$ | 00:35:03 to 00:37:47 | 'So duplication [of data] can be good?' | Yes |
| $R_{11}$ | 00:37:47 to 00:43:10 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $R_{12}$ | 00:43:10 to 00:46:49 | **6) [Ch3-S3.4.5-T2]** | Yes |
| $R_{13}$ | 00:46:49 to 00:47:37 | 'What about *um* quality assurance information – is that an important aspect of the meta- – provenance metadata as well?' | Yes |
| $R_{14}$ | 00:47:37 to 00:50:02 | **7) [Ch3-S3.4.5-T2]** | No |
| $R_{15}$ | 00:50:02 to 00:58:16 | 'What about copyright and data licensing? | Yes |
| *Interview stopped for a couple of minutes. First half of interview completed at 00:58:16* | | | |
| $R_{16}$ | 00:00:00 to 00:07:02 | **8) [Ch3-S3.4.5-T2]** | No |
| $R_{17}$ | 00:07:02 to 00:11:20 | **9) eCrystals and LabTrove [Ch3-S3.4.5-T2]** | No |
| $R_{18}$ | 00:11:20 to 00:14:00 | **10) [Ch3-S3.4.5-T2]** | No |
| *Second half of Interview completed at 00:58:16 + 00:14:00* | | | |

# B.3 Chapter 5: FLLOC and SPLLOC case study

## B.3.1 Mr D. record of interview questions

| **Mr D.** Record of Interview Questions | | | |
|---|---|---|---|
| **Interview Section:** | **Question Duration:** | **Question:** | **Directly Referenced in thesis?** |
| $D_1$ | 00:00:00 to 00:01:08 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $D_2$ | 00:01:08 to 00:03:26 | **2) i) [Ch3-S3.4.5-T2]** | No |
| $D_3$ | 00:03:26 to 00:07:36 | **2) ii) [Ch3-S3.4.5-T2]** | No |
| $D_4$ | 00:07:36 to 00:09:36 | **3) [Ch3-S3.4.5-T2]** | Yes |
| $D_5$ | 00:09:36 to 00:13:01 | '*Um* do you think a lot of researchers are quite technically aware? They can, sort of, they – they can use tools quite easily – and things are quite user-friendly? Or do you think they rely on technical support?' | No |
| $D_6$ | 00:13:01 to 00:15:20 | '*Um* what do you think of 'persistent identifiers?' Do you think it's right that – it's good that they're not necessarily going to break? But, do you think it's right that you rely on just one company?' [**Mr D.** responds: 'Well tell me what you mean by 'persistent identifier'?' **LG** responds: 'So, *um* I think 'cause DOIs have become quite popular with a lot of *um* services, sort of publications – even I think in eCrystals repository they're starting to link them through the DataPool project. They decided they would trial them through that. *Um*, because obviously when people change websites around sometimes the links can break. *Um* but from other research I've been looking at, some people might argue that: 'is it right that we should be giving all this information to one organisation?'] | Yes |
| $D_7$ | 00:15:20 to 00:15:59 | 'Do you think that's sort of the '.ac.uk' – 'soton.ac.uk' sort of helps with trust as well, because you're sort of branding research projects – projects on the Web?' | Yes |
| $D_8$ | 00:15:59 to 00:18:28 | **4) [Ch3-S3.4.5-T2]** | Yes |
| $D_9$ | 00:18:28 to 00:20:07 | '*Er* why do you think academic [research] data's suddenly become – it's suddenly become sort of more important than before? Whereas like publications were the main focus, but now people are really starting to think: 'oh, we want to mandate that we have open data and …' | Yes |
| $D_{10}$ | 00:20:07 to 00:22:08 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $D_{11}$ | 00:22:08 to 00:25:10 | 'What do you think about Creative Commons and open licensing systems like that?' | No |

| | | | |
|---|---|---|---|
| $D_{12}$ | 00:25:10 to 00:26:08 | **6)** [**Mr D.** asks: 'Provenance metadata … Provenance metadata as distinct to descriptive metadata?' **LG** responds: 'So provenance metadata, sort of, I've defined it as information about like who created [the datum] it, permissions, formats – all that kind of stuff.'] **[Ch3-S3.4.5-T2]** | Yes |
| $D_{13}$ | 00:26:08 to 00:30:33 | '*Um* sort of, what's come up before is – people have concerns over how easy is it to de-link provenance data from a dataset, and how easy is it to de-link the licensing information as well – if someone really wanted to?' | No |
| $D_{14}$ | 00:30:33 to 00:34:38 | '*Um* what do you think about how provenance should be displayed? Do you think it should be [via] Dublin Core metadata standards, or do you think …?' | Yes |
| $D_{15}$ | 00:34:38 to 00:36:01 | '*Um* do you think sometimes then just having a PDF – where you've just written up your own provenance – is enough? Is that going the opposite direction …?' | No |
| $D_{16}$ | 00:36:01 to 00:40:28 | 'Is it better for longevity because, you know, new languages [formats] come and go that *um* documents are just sort of maybe just given an xml file and everything's just put – or a .txt file? Just … But then it's very basic.' | No |
| $D_{17}$ | 00:40:28 to 00:40:57 | 'So, what do you think about linked data? How has it had an impact on academic research data so far?' | No |
| $D_{18}$ | 00:40:57 to 00:43:01 | 'But do you think [linked data] it could be useful [for academic research data re-usage] in the future?' | No |
| $D_{19}$ | 00:43:01 to 00:43:20 | '*Um* I've only looked at it briefly, but *um* I was looking at [Professor] Luc Moreau's work – his team's work on the W3C for provenance language – I was just wondering how that would apply to research data? | No |
| $D_{20}$ | 00:43:20 to 00:47:53 | **7)** [**Mr D.** asks: 'External issues … I don't know, I mean – what's your definition of 'external'?' **LG** responds: 'I think when I was *um* talking about external issues, sort of external to the … the … the … [external] legal issues like *um* copyright and things that I was focusing on.'] **[Ch3-S3.4.5-T2]** | Yes |
| $D_{21}$ | 00:47:53 to 00:50:57 | '*Um* what do you think about sensitive data, sort of, for example interviews where people might not consent to it being – they don't want the sound recordings open – but do you think things should be made discoverable? So, you should put on the Web somewhere that: 'I've conducted these [interviews] and this is where you can find my research, but the raw – you can't access these raw sound recordings'?' | Yes |
| $D_{22}$ | 00:50:57 to 00:51:30 | '*Um* do you think that's been an issue with academic publishing and peer review that they sometimes don't check all the underlying data?' | No |
| $D_{23}$ | 00:51:30 to | **8) [Ch3-S3.4.5-T2]** | Yes |

| | | | |
|---|---|---|---|
| | 00:54:17 | | |
| $D_{24}$ | 00:54:17 to 00:56:15 | 'Do you think computer science might become actually more important to school education, as now it's so integral to every part of everyone's lives? […] Because I know they [UK government] spoke about putting it on the education agenda.' | No |
| $D_{25}$ | 00:56:15 to 00:57:56 | '*Um* just one other thing about on the Semantic Web – do you think it's actually very user-friendly at the moment? [For example] Somebody, I suppose maybe, I know in the law building has no knowledge of these things – could actually, sort of, [think] I want to go and make my – dataset – linked data – and use these things?' | No |
| $D_{26}$ | 00:57:56 to 00:59:28 | **9) [Ch3-S3.4.5-T2]** | No |
| $D_{27}$ | 00:59:28 to 00:59:38 | **10) [Ch3-S3.4.5-T2]** | No |
| *Interview completed at 00:59:38* | | | |

## B.3.2 Mr K. record of interview questions

| **Dr K.** Record of Interview Questions | | | |
|---|---|---|---|
| Interview Section: | Question Duration: | Question: | Directly Referenced in thesis? |
| $K_1$ | 00:00:00 to 00:01:54 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $K_2$ | 00:01:54 to 00:04:39 | **2) i)** [**Dr K.** asks: '*Um* can you just maybe flesh out what, what that question means a bit more in terms of what does … *er* 'academic research data re-usability' … what does that mean?' **LG** responds: 'So I'm looking at *um* sort of on the Web – how the Web's facilitating *um* academic data re-usability. So, *um* has it improved access to data or sign-posting of data? Are there any sort of data repositories or any way we can access *um* these types of data on the Web?'] **[Ch3-S3.4.5-T2]** | No |
| $K_3$ | 00:04:39 to 00:06:29 | **2) ii) [Ch3-S3.4.5-T2]** | Yes |
| $K_4$ | 00:06:29 to 00:09:47 | **3) [Ch3-S3.4.5-T2]** | Yes |
| $K_5$ | 00:09:47 to 00:10:24 | '*Um* do you think that all oral histories data should remain or – is it part of the process that some will get deleted due to confidentiality?' | Yes |
| $K_6$ | 00:10:24 to 00:12:00 | 'Do you think it might put some people off *um* being interview participants if they're told that: 'I'm going to make your sound recording available to all on the Web'? | Yes |
| $K_7$ | 00:12:00 to 00:13:23 | **4) [Ch3-S3.4.5-T2]** | No |
| $K_8$ | 00:13:23 to 00:15:07 | 'Well I actually think for this question I might frame it the other way and look at *um* when it comes to publishers how do they quality assure *um* oral histories data? Do they usually take copies? Or do they just trust in the researcher and that institution's ethical process?' | Yes |
| $K_9$ | 00:15:07 to 00:15:30 | '*Um* when it comes to funding bodies do they mandate that this [oral histories] data should be made openly available?' | No |
| $K_{10}$ | 00:15:30 to 00:17:32 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $K_{11}$ | 00:17:32 to 00:18:39 | '*Um* have you ever re-used anyone else's [oral histories] dataset? And did you have to get permission – sort of copyright clearance [and] things like that?' | No |
| $K_{12}$ | 00:18:39 to 00:19:40 | '*Um* so do you think that institutional repositories – like ePrints [Soton] – should really support researchers to deposit data of this type [oral histories]? And do you think it should end up in institutional repositories where possible?' | No |
| $K_{13}$ | 00:19:40 to 00:20:09 | '*Um* are there any sort of registries of oral histories data so you can, sort of, see even though they're not publically available or maybe they've been deleted, you can see that they've existed – anywhere on the Web?' | No |
| $K_{14}$ | 00:20:09 to 00:20:59 | 'Is there anything – a centralised place in the UK [for oral histories data]?' | No |
| $K_{15}$ | 00:20:59 to 00:22:08 | **6) [Ch3-S3.4.5-T2]** | No |
| $K_{16}$ | 00:22:08 to 00:22:40 | '*Um* do you think it's difficult to re-use oral histories data, as the data is so closely connected to the research | No |

| | | | |
|---|---|---|---|
| | | question?' | |
| $K_{17}$ | 00:22:40 to 00:24:28 | **7) [Ch3-S3.4.5-T2]** | No |
| $K_{18}$ | 00:24:28 to 00:25:25 | 'Yeah, more yeah, exploitation and that for commercial use I suppose. I suppose if you are using [oral histories data] to write a book [and] that creates revenue as well … it's more indirect?' | No |
| $K_{19}$ | 00:25:25 to 00:26:48 | '*Um* do you think it's the University's responsibility to support … *um* … researchers undertaking oral histories research – to provide them [with] good ethics guidelines and … *um* … good codes of conduct and 'support documentation'?' | Yes |
| $K_{20}$ | 00:26:48 to 00:29:17 | '*Um* … what … well, well what impact do you think open access has had [on oral histories research]?' | Yes |
| $K_{21}$ | 00:29:17 to 00:31:10 | **8) [Ch3-S3.4.5-T2]** | No |
| $K_{22}$ | 00:31:10 to 00:31:43 | '*Um* … do many people in oral histories work in collaboration with *um* computer scientists or anything? Sort of help them make digital methodologies or …? | Yes |
| $K_{23}$ | 00:31:43 to 00:36:04. | **9) i)** [**LG** adds: 'I was thinking sort of within the sciences – some of the data – it's very fast-paced and it's non-personal – it's from experiments. It can be put out there. You haven't got the sensitive issues. So you could – it's more transparent, so you could sort of see the process and when it was created, and what happened, all the way along. Whereas, oral histories, maybe some of the … the real names and things are hidden so who can sort of quality assess that, as transparently?'] **[Ch3-S3.4.5-T2]** | No |
| $K_{24}$ | 00:36:04 to 00:39:05 | **9) ii) iii) [Ch3-S3.4.5-T2]** | Yes |
| $K_{25}$ | 00:39:05 to 00:41:07 | '*Um* I know for my own PhD, my supervisors and examiners will be allowed to listen to the sound recordings, but it won't go any further. When you're beyond PhD, as an academic, do you have other peers listening in? Are they allowed to listen into this data even though it can't be publically available? Or is that down to the particular interviewer? It might just be the interviewer who can listen to the sound recordings?' | No |
| $K_{26}$ | 00:41:07 to 00:43:13 | '*Um* what do you think about transcription software?' | No |
| $K_{27}$ | 00:43:13 to 00:43:27 | 'Has that ever happened where they've [oral historians] have modified raw data the sound recordings? Maybe got someone else to read it out in a different accent to try and hide the true person behind it?' | No |
| $K_{28}$ | 00:43:27 to 00:43:42 | **10) [Ch3-S3.4.5-T2]** | No |
| *Interview completed at 00:43:42* | | | |

### B.3.3 Miss L. record of interview questions

| **Miss L.** Record of Interview Questions | | | |
|---|---|---|---|
| Interview Section: | Question Duration: | Question: | Directly Referenced in thesis? |
| $L_1$ | 00:00:00 to 00:01:09 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $L_2$ | 00:01:09 to 00:06:55 | **2) i) [Ch3-S3.4.5-T2]** | No |
| $L_3$ | 00:06:55 to 00:12:12 | '*Um* how does this link up to existing repositories in the University, *er* such as eCrystals, who are already publishing data?' | No |
| $L_4$ | 00:12:12 to 00:16:10 | **2) ii) [Ch3-S3.4.5-T2]** | Yes |
| $L_5$ | 00:16:10 to 00:19:21 | **3) [Ch3-S3.4.5-T2]** | Yes |
| $L_6$ | 00:19:21 to 00:22:33 | '*Um* what about the potential for orphan data in the future? How can you maintain relationships with the people who, perhaps, created datasets – but maybe they changed jobs, or if it's two-hundred years' time people aren't around anymore?' | No |
| $L_7$ | 00:22:33 to 00:26:36 | **4) [Ch3-S3.4.5-T2]** | Yes |
| $L_8$ | 00:26:36 to 00:31:20 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $L_9$ | 00:31:20 to 00:35:30 | 'How important is data licensing? Does it make the legal issues more accessible to people, perhaps, who don't have or have a limited knowledge of the law of re-use?' | No |
| $L_{10}$ | 00:35:30 to 00:40:36 | '*Um* has it been difficult with the *um* DataPool project, to actually change internal academic practices? *Um* have some disciplines more – been more resilient to change?' | No |
| $L_{11}$ | 00:40:36 to 00:44:32 | **6) [Ch3-S3.4.5-T2]** | Yes |
| $L_{12}$ | 00:44:32 to 00:48:24 | **7) [Ch3-S3.4.5-T2]** | Yes |
| $L_{13}$ | 00:48:24 to 00:53:28 | **8) [Ch3-S3.4.5-T2]** | No |
| $L_{14}$ | 00:53:28 to 00:59:32 | **9) [Miss L.** asks: '*Um* … When you say: 'personal' – do you mean personal to the individual that you're recording or personal to the interviewer?' **LG** responds: 'To the individual that you're recording.'] **[Ch3-S3.4.5-T2]** | Yes |
| $L_{15}$ | 00:59:32 to 01:02:14 | '*Um* do you think data destruction is a part of *erm* data re-use as well?' | No |
| $L_{16}$ | 01:02:14 to 01:03:33 | **10) [Ch3-S3.4.5-T2]** | No |
| $L_{17}$ | 01:03:33 to 01:05:49 | '*Um* obviously 'cause copyright doesn't subsist in all datasets, but I think data licensing – sometimes you – you licence all data for all re-use even though maybe copyright doesn't exist – subsist anyway …' | No |
| *Interview completed at 01:05:49* | | | |

## B.3.4 Mr M. record of interview questions

| **Mr M.** Record of Interview Questions | | | |
|---|---|---|---|
| Interview Section: | Question Duration: | Question: | Directly Referenced in thesis? |
| $M_1$ | 00:00:00 to 00:02:59 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $M_2$ | 00:02:59 to 00:05:11. | **2) [Ch3-S3.4.5-T2]** | Yes |
| $M_3$ | 00:05:11 to 00:11:09 | **3) [Ch3-S3.4.5-T2]** | Yes |
| $M_4$ | 00:11:09 to 00:13:38 | '*Um* what about individual researcher's liability [for misuse of academic research data]?' | No |
| $M_5$ | 00:13:38 to 00:17:48 | **4) [Ch3-S3.4.5-T2]** | No |
| $M_6$ | 00:17:48 to 00:24:37 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $M_7$ | 00:24:37 to 00:28:55 | '*Um* do you think *um* academics who maybe don't have a legal background … well, copyright sometimes it can be a grey area whether it subsists in data or not, and academics maybe don't understand whether they do need a data licence or … why there's contract law and copyright law?' | Yes |
| $M_8$ | 00:28:55 to 00:31:58. | '*Um* … What do you think the place is for moral rights in academic research data re-usability?' | No |
| $M_9$ | 00:31:58 to 00:32:50 | **6) [Ch3-S3.4.5-T2]** | Yes |
| $M_{10}$ | 00:32:50 to 00:37:16 | **7) [Ch3-S3.4.5-T2]** | Yes |
| $M_{11}$ | 00:37:16 to 00:41:20. | **8) [Ch3-S3.4.5-T2]** | No |
| $M_{12}$ | 00:41:20 to 00:42:59. | '*Um* … Do you think that *er* … well, repositories, such as eCrystals, and academic disciplines, such as oral narratives, they have sort of a commonality of problems when it comes to data re-use? Or do you think they're really – really separate … separated?' | No |
| $M_{13}$ | 00:42:59 to 00:45:12. | 'You mentioned ISO earlier – do you think they have a big role with academic data re-usability?' […] Because I don't know that they have some metadata standards that are used – obviously do you have to pay for them personally as a researcher at a University?' | No |
| $M_{14}$ | 00:45:12 to 00:48:54 | '*Um* you know the policies of *er* sort of best practice and things – do you sort of *um* … who do you ask? Who do you involve? Do you have working groups?' | No |
| $M_{15}$ | 00:48:54 to 00:50:57 | '*Um* what is the worst problem that you've come across with copyright clearance?' | No |
| **Question 9) was not asked directly as the points were covered throughout the interview. Question 10) was not asked as there were no further points to be discussed.** | | | |
| *Interview completed at 00:50:57* | | | |

## B.3.5 Dr O. record of interview questions

| **Dr O.** Record of Interview Questions | | | |
|---|---|---|---|
| Interview Section: | Question Duration: | Question: | Directly Referenced in thesis? |
| $O_1$ | 00:00:00 to 00:00:58 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $O_2$ | 00:00:58 to 00:04:54 | **2) i) [Ch3-S3.4.5-T2]** | No |
| $O_3$ | 00:04:54 to 00:05:48 | **2) ii) [Ch3-S3.4.5-T2]** | Yes |
| $O_4$ | 00:05:48 to 00:07:43 | '*Um* how popular are these [CHILDES, FLLOC and SPLLOC] databases?' | Yes |
| $O_5$ | 00:07:43 to 00:08:22 | '*Um* are there any examples – what examples are there of re-use – re-uses? Do you have *um* teachers at school using these [CHILDES, FLLOC and SPLLOC] datasets or …?' | Yes |
| $O_6$ | 00:08:22 to 00:12:03 | **3) [Ch3-S3.4.5-T2]** | Yes |
| $O_7$ | 00:12:03 to 00:13:49 | **4) [Ch3-S3.4.5-T2]** | Yes |
| $O_8$ | 00:13:49 to 00:14:34 | '*Um* traditionally with the academic journal publishing – did they want to see *um* raw datasets? So, they wanted to see the actual … they wanted to have copies of the sound recording to quality assess journal articles?' | Yes |
| $O_9$ | 00:14:34 to 00:15:26 | **5) [Ch3-S3.4.5-T2]** | Yes |
| $O_{10}$ | 00:15:26 to 00:16:06 | 'What about licensing systems for the data? I think I've seen it uses – *um* I think CHILDES – was the Creative Commons unported licence … I'm just having a look [**LG** searches through notes]… at the ground rules…' | Yes |
| $O_{11}$ | 00:16:06 to 00:16:30 | '*Um* have you sort of…do … have you, sort of, worked with Legal Services at the University – about any of the legal issues?' | Yes |
| $O_{12}$ | 00:16:30 to 00:17:04 | '*Um* just a final thing, I suppose, on the legal issues. *Um* when it comes to copyright law I'm – it will vest in your verbatim transcripts and things – that will vest in the author – *um* has that been an issue with re-use?' | No |
| $O_{13}$ | 00:17:04 to 00:17:19 | 'So has, sort of, all the legal issues been sorted out by CHILDES, sort of, in the [United] States? – They've sort of done all the kind of infrastructure? Okay, so you don't have specific licences for different – each dataset.' | No |
| $O_{14}$ | 00:17:19 to 00:18:53 | **6)** [**Dr O.** asks: 'Do that – When you mean 'metadata' that's information about the …' **LG** responds: '…data – about who created the data, the date it's created, the format …'] **[Ch3-S3.4.5-T2]** | Yes |
| $O_{15}$ | 00:18:53 to 00:19:04 | '*Um* is the metadata quite standardised across all the [FLLOC and SPLLOC] projects then?' | Yes |
| $O_{16}$ | 00:19:04 to 00:19:35. | 'And *um* how does … how do the projects link up with ePrints [Soton] at the University? Is [sic] there links through articles …' | No |
| $O_{17}$ | 00:19:35 to 00:22:19 | **7) [Ch3-S3.4.5-T2]** | Yes |
| $O_{18}$ | 00:22:19 to 00:23:03 | 'Do you think, in general, people wanted to participate [in FLLOC and SPLLOC]?' | No |
| $O_{19}$ | 00:23:03 to 00:23:18 | '*Um* what did you do with the *um* [FLLOC and SPLLOC] data about their names and the school names? Did you – Have you kept them as lists somewhere | Yes |

| | | | |
|---|---|---|---|
| | | separate to the data?' | |
| $O_{20}$ | 00:23:18 to 00:23:41 | 'Is that [keeping FLLOC and SPLLOC lists of names of participants and schools] just for future reference if someone comes to you – and quality assurance, so people can check back?' | Yes |
| $O_{21}$ | 00:23:41 to 00:23:49 | 'Did you alter the voices [in the FLLOC and SPLLOC sound files] at all?' | No |
| $O_{22}$ | 00:23:49 to 00:26:17 | **8) [Ch3-S3.4.5-T2]** | Yes |
| $O_{23}$ | 00:26:17 to 00:27:12 | '*Um* what do you think about the open access and open data movements? | Yes |
| $O_{24}$ | 00:27:12 to 00:28:32 | 'How much *um* data is open on the FLLOC [and SPLLOC] websites'? | No |
| $O_{25}$ | 00:28:32 to 00:29:10 | 'What about the Data Protection Act [1998]? [**Dr O.** asks: 'In terms of access to [FLLOC and SPLLOC data] it?' **LG** responds: 'And *um*, sort of, you're looking after this personal data [lists of names of participants] that's been anonymised?' | No |
| $O_{26}$ | 00:29:10 to 00:29:48 | '*Um* how has the Web really impacted [on] this area of [second language acquisition] research?' | No |
| $O_{27}$ | 00:29:48 to 00:31:10 | **9) i) [Ch3-S3.4.5-T2]** | No |
| $O_{28}$ | 00:31:10 to 00:31:33 | 'Do you need CRB [Criminal Records Bureau, now known as Disclosure and Barring Service (DBS)] checks to carry out this research?' | Yes |
| $O_{29}$ | 00:31:33 to 00:31:46 | '*Um* how important is attribution to people working within this [second language acquisition research] area?' [**Dr O.** asks: '*Er* … In terms of saying who's collecting the data?' **LG** responds: 'Yeah'.] | No |
| $O_{30}$ | 00:31:46 to 00:32:09 | '*Um* do you ever worry about a de-linking between this … this [provenance] data just being deleted and then someone just re-using your data without attributing you or acknowledging you?' | No |
| $O_{31}$ | 00:32:09 to 00:32:42 | 'Is *um* is this community – [second] language acquisition research – quite a small community – do you, sort of, know a lot of people?' | No |
| $O_{32}$ | 00:32:42 to 00:34:27 | 'If *um* there were more, like a – a bigger research community, say [researchers] didn't know the majority of people [in their research area] – do you think it would affect the trust and the attribution issues?' | No |
| $O_{33}$ | 00:34:27 to 00:34:36 | '*Um* would you say historically in this [second language acquisition] discipline there has always been this culture of sharing?' | Yes |
| $O_{34}$ | 00:34:36 to 00:35:28 | 'Because obviously in some other discipline people want to 'hug' their data …' | No |
| $O_{35}$ | 00:35:28 to 00:36:22 | '*Er* what do see the future of this area being? Do you think it might diversify or …?' | Yes |
| $O_{36}$ | 00:36:22 to 00:37:46 | '*Um* is this research community quite interdisciplinary, because I – I was reading sort of some people might use certain data for *um* computational linguistics …' | Yes |
| $O_{37}$ | 00:37:46 to 00:38:10 | **10) [Ch3-S3.4.5-T2]** | No |
| **Question 9) ii) and iii) were not asked directly as the points were covered throughout the interview.** | | | |
| *Interview completed at 00:38:10* | | | |

## B.3.6 Dr P. record of interview questions

| **Dr P.** Record of Interview Questions | | | |
|---|---|---|---|
| Interview Section: | Question Duration: | Question: | Directly Referenced in thesis? |
| $P_1$ | 00:00:00 to 00:01:10 | **1) [Ch3-S3.4.5-T2]** | Yes |
| $P_2$ | 00:01:10 to 00:02:57 | **2) i) [Ch3-S3.4.5-T2]** | No |
| $P_3$ | 00:02:57 to 00:03:51 | 'How has this [LANGSNAP] built on, sort of, prior research in schools that you [part of the LANGSNAP research team] did with the former projects [FLLOC and SPLLOC]?' | No |
| $P_4$ | 00:03:51 to 00:07:28 | **2) ii) [Dr P.** asks: 'Yeah, okay – could you split that up? So the yeah …so could you split that up?' **LG** responds: '*Um* so … sort of what software are you using for your database?'] **[Ch3-S3.4.5-T2]** | Yes |
| $P_5$ | 00:07:28 to 00:08:30 | 'Have you had any major problems with any of these [software] tools [i.e. CHILDES, Praat, ELAN and iSurvey]?' | Yes |
| $P_6$ | 00:08:30 to 00:09:43 | *'Um* … Does the [second language acquisition] research community – do you have a lot of, sort of impact on the kind of software that's used? Do you work with computer scientists and programmers?' | Yes |
| $P_7$ | 00:09:43 to 00:11:51 | **3)** [**Dr P.** asks: 'So *er* … So when you say 'quality assurance' can you [explain that]…?' **LG** responds: '*Um* so, like the accuracy of the data, the completeness, the timeliness – so that others re-using it could know that it's reliable?'] **[Ch3-S3.4.5-T2]** | Yes |
| $P_8$ | 00:11:51 to 00:13:55 | '*Um* … Did you find that a lot of people *um* on their year abroad wanted to actually give you their data, sort of, make their interviews openly available on the Web?' | Yes |
| $P_9$ | 00:13:55 to 00:16:14 | **4) [Ch3-S3.4.5-T2]** | No |
| $P_{10}$ | 00:16:14 to 00:19:16 | '*Um* … Do you think the traditional publication process where you publish articles about this research where maybe you wouldn't provide the raw *um* sound recordings – has the Web really enabled you now to, sort of, in a way bypass that and just say: 'here's our database of sound r[ecordings]'? Or, were things like that happening before?' | Yes |
| $P_{11}$ | 00:19:16 to 00:20:21 | 'What do you also think about the open data movement? – So making the actual underlying data of publications openly accessible?' **Dr P** asks: 'I think it's a… so say that again'. **LG** responds: 'So *um* the open data movement. So making the underlying data within traditional publications accessible? Yeah, do you think sort of … I don't know in the past if *um* … you could easily get hold of databases from – if you read, I don't know *um* a past project from twenty years or so ago – could you get hold of that data quite easily [today]?' | No |
| $P_{12}$ | 00:20:21 to 00:22:21. | **5) [Ch3-S3.4.5-T2]** | Yes |
| $P_{13}$ | 00:22:21to 00:23:04 | '*Um* … What about data licensing and copyright – and other intellectual property rights? Has that come up in this [LANGSNAP] project?' | No |
| $P_{14}$ | 00:23:04 to 00:23:12 | '*Um* so do you [LANGSNAP] have any licensing agreements in place? [**Dr P.** says: 'No'.] Or do you | No |

| | | | |
|---|---|---|---|
| | | have terms and conditions?' | |
| $P_{15}$ | 00:23:12 to 00:24:34 | '*Um* … Have you [LANGSNAP] worked within the … any … any like legal professionals in the University [of Southampton] like Legal Services or Research and Innovation Services?' | Yes |
| $P_{16}$ | 00:24:34 to 00:25:06 | '*Er* what do you think of open licensing systems like Creative Commons and things like that? Do you come across – in your wider research – anything of that kind of nature?' | No |
| $P_{17}$ | 00:25:06 to 00:25:45 | 'Do you have … do you face any copyright issues in your daily work or not really?' | Yes |
| $P_{18}$ | 00:25:45 to 00:28:08 | **6)** [**Dr P.** asks: '[Provenance metadata] Which is?' **LG** responds: 'So it's sort of the information about 'the data about the data' so … where it came from, the dates, the times it's collected … *um* you know, what formats it's in and that kind of thing.'] **[Ch3-S3.4.5-T2]** | Yes |
| $P_{19}$ | 00:28:08 to 00:31:17 | **7) [Ch3-S3.4.5-T2]** | Yes |
| $P_{20}$ | 00:31:17 to 00:31:49 | '*Um* … because surely if you think of statistical support there must be people in the social sciences faculty or even maths that could help? But is it just because – I don't know, the culture of the university it's not very interdisciplinary or cross-faculty – that people don't get together?' | No |
| $P_{21}$ | 00:31:49 to 00:32:13 | 'So, sort of, when you're making a project bid you can't, sort of, put a pot aside as like a contingency fund can you? You have to say that: 'this is what we're doing it for' – and that's a problem?' | No |
| $P_{22}$ | 00:32:13 to 00:32:23 | 'Is there any way you can update bids? Or when it's been accepted that's it?' | No |
| $P_{23}$ | 00:32:23 to 00:33:47 | 'Do you think as projects become more real-time – if, sort of, that you've got a website and you're updating it, and funding councils can see what you're doing – so that you're showing results – than really a funding bid in a way should become more real-time?' **Dr P.** asks: 'What do you mean?' **LG** responds: 'Sort of if they can see that you're having certain impacts and that … then … you then down the line think: 'oh I need some statistical support and we need an extra pot of money for that', can … in a way the funding council could actually see that … what you've done. And maybe think: 'oh they're having a valuable impact' and maybe we can re-assess [the funding bid]?' | No |
| $P_{24}$ | 00:33:47 to 00:35:36 | **8) [Ch3-S3.4.5-T2]** | No |
| $P_{25}$ | 00:35:36 to 00:37:01 | '*Um* … Do you see any other impacts, sort of, within that – the software and systems – or do you think CHILDES will be the main sort of hub? For all these… I suppose CHILDES is it sort of like the … where it sort of records a lot of projects, because obviously it takes a copy. So really it is sort of acting as this sort of centralised database, and then linking to, let's say, prior projects you've done with LANGSNAP, and then other projects around the world as well? Yeah.' **Dr P.** responds: 'That what it does, yeah'. **LG** asks: 'Yeah. So do you think people will be maybe … when they're trying to search for data will be going to that [CHILDES] database first and then …?' | Yes |
| $P_{26}$ | 00:37:01 to 00:39:02 | **9) i) [LG** adds: So in other words, sort of like, if we're looking at eCrystals within chemistry – it's very | Yes |

| | | | |
|---|---|---|---|
| | | structured, formulaic data, it's … it's not got that human element. *Um* so, sort of what are the different challenges that second language acquisition research has?'] | |
| $P_{27}$ | 00:39:02 to 00:39:57 | '*Um* what about, sort of … Are any of the interviewers sensitive about their data, about the questions they've asked as well? *Um* … Do they remain anonymous?' [**…**] 'Even the interviewers as well?' | No |
| $P_{28}$ | 00:39:57 to 00:41:29 | **9) ii) iii) [Ch3-S3.4.5-T2]** | Yes |
| $P_{29}$ | 00:41:29 to 00:43:13 | '*Um* are you sort of – do you see any problems with relying on just one kind of piece of software as CHILDES – because it's … it's Carnegie Mellon University in the States? *Um* do you worry that they're sort of making the standards or do they ask the research community … the global research community what they'd like to see?' | Yes |
| $P_{30}$ | 00:43:13 to 00:43:34 | 'Without CHILDES – if it didn't exist – what do you think would've happened to these projects?' | No |
| $P_{31}$ | 00:43:34 to 00:44:23 | 'Is … is 'Esmeralda' sort of an open sort of software as well?' | Yes |
| $P_{32}$ | 00:44:23 to 00:45:57 | **10)** [**Dr P.** asks: '*Erm* just so, so what you're interested is the role of the Web is it?' **LG** responds: 'Yeah, *erm* sort of different academic research data re-use systems on the Web, and sort of, in different … in three different areas, so yeah …'] **[Ch3-S3.4.5-T2]** | No |
| $P_{33}$ | 00:45:57 to 00:46:32 | '*Um* do you have a historic backlog of data before the sort of digital age – before the Web? That's sort of maybe stuck in just *um*, I don't know, on a floppy disk or something somewhere, or any sound recordings?' | Yes |
| $P_{34}$ | 00:46:32 to 00:47:10 | 'Has there been any data loss, because of [digitisation]?' | No |
| $P_{35}$ | 00:47:10 to 00:48:11 | 'What do you think about data destruction methods – some researchers maybe have interviews that their participants aren't happy with, sort of, being made available, so they might get destroyed – do you think that's good for transparency or do you think it … it's needed sometimes for ethical issues?' | No |
| $P_{36}$ | 00:48:11 to 00:48:17 | 'Is there anything else?' | No |
| *Interview completed at 00:48:17* | | | |

# Glossary

This thesis forms part of interdisciplinary, web science research practices. It is produced for an interdisciplinary audience and is supervised by three faculties at the University of Southampton: the Faculty of Humanities, the Faculty of Business, Law and Art; and the Faculty of Physical Sciences and Engineering. To avoid any ambiguity where certain terms are discipline specific or have distinct meaning within different disciplines, the following definitions are understood throughout:

**Academic research data** – any qualitative or quantitative data that have the potential for (re)use, either partially or completely that are produced in the course of academic research within the sciences, social sciences and humanities

**Ahistoricism** – the failure to acknowledge the longer heritage of knowledge transfer, and its historical and contextual development across pre-digital and digital ages

**Attribution stacking** – managing multiple types of attribution statements pertaining to various academic research datasets

**Custodian** – individual(s) that are authorised to safeguard academic research data on behalf of data originators

**Data** – all qualitative (non-numerical) and quantitative (numerical) materials, which include: facts, observations, measurements, statistics, figures, lists, sound recordings, verbatim transcripts, bibliographies, memoirs, case law and statutes; materials obtained for interpretation and analysis to produce information, knowledge and/or wisdom

**Data originator** – an individual who gathers a specific academic research dataset

**Data/text mining** – the analytical use of algorithms to find patterns across a collection of multiple datasets and/or texts

**Diffractometer** – specialist laboratory equipment used by crystallographers

**DOI minting** – the process of creating digital object identifiers for academic research datasets

**Knowledge transfer** – sharing high quality information within a user community

**Licensing stacking** – managing multiple types of data licence agreements pertaining to various academic research datasets

**Machine-readable** – a machine can parse the data and provide an unambiguous structure (e.g. HTML and XML are machine-readable), but cannot provide any inherent meaning in the structure (i.e. tag *y* means *x*)

**Machine-understandablity** – a machine cannot only provide an unambiguous structure, but its inherent meaning (e.g. RDF is machine-understandable)

**Metadata** – data about an academic research dataset

**Meta-metadata** – provenance information concerning the provenance metadata attached to an academic research dataset

**Moral rights** – [legal term] personal, legal rights (not about morality); four moral rights are conferred by the Copyright, Designs and Patents Act (CDPA) 1988, section 79 to section 89 in the UK: 1) right of paternity; 2) the right to integrity; 3) the right to object to false attribution; and 4) the right to privacy in photographs and films

**Originality** – [legal term] there is no copyright in ideas, but in their tangible expression; in UK copyright law, for copyright to subsist in literary, dramatic, musical and artistic works there is a requirement of legal originality; the standard of originality is low – innovation is not required, but works must derive from the author(s)

**Provenance** – a digitally accessible record pertaining to a specific academic research dataset, including information about its origins, development, versions, legal, technological and socio-cultural frameworks

**Qualitative data –** non-numerical data – often textual materials; for example case law, bibliographies, interview sound recordings and verbatim transcriptions

**Quantitative data** – numerical data; for example, statistics, numerical measurements and graphs

**Research user** – an individual who requires access to high quality academic research data

**Re-usage** – research users accessing, evaluating, analysing and interpreting academic research data collected by third parties

**Usage** – data originators accessing, evaluating, analysing and interpreting the academic research data that they collected

**User community** – a group of research users who require access to (similar) types of high quality academic research data

**Web of linked data (Semantic Web)** – machine-understandable data published on the Web resulting in greater interoperability and integration

**Web of linked documents** – machine-readable documents published on the Web connected through hyperlinks

# Bibliography

**Primary case study materials:**

*eCrystals Website* <http://ecrystals.chem.soton.ac.uk/> [accessed 9 August 2015]

*French Learner Language Oral Corpora (FLLOC) Website* <http://www.flloc.soton.ac.uk/> [accessed 9 August 2015]

*LabTrove Website* <http://www.labtrove.org/> [accessed 9 August 2015]

*Languages and Social Networks Abroad Project (LANGSNAP) Website* <http://langsnap.soton.ac.uk/> [accessed 9 August 2015]

*Spanish Learner Language Oral Corpora (SPLLOC) Website* <http://www.splloc.soton.ac.uk/> [accessed 9 August 2015]

*The Archive for Marine Species and Habitats Data (DASSH) Website* <http://www.dassh.ac.uk/> [accessed 9 August 2015]

*The British Oceanographic Data Centre (BODC) Website* <http://www.bodc.ac.uk/> [accessed 9 August 2015]

*The Marine Environmental Data and Information Network (MEDIN Website)* <http://www.oceannet.org/> [accessed 9 August 2015]

*UK Hydrographic Office (UKHO) Website* < http://www.ukho.gov.uk/> [accessed 9 August 2015]

**Other source materials:**

'About the Open Government Licence', *The National Archives Website* <http://www.nationalarchives.gov.uk/information-management/re-using-public-sector-information/licensing-for-re-use/what-ogl-covers/> [accessed 9 August 2015]

'Academic Licence for the use of data supplied to the BODC: 1 km x 1 km gridded bathymetry for Irish Sea, Celtic Sea and North Channel', *The British Oceanographic Data Centre (BODC) Website* <http://www.bodc.ac.uk/products/external_products/celtic_seas/documents/licence.pdf> [accessed 9 August 2015]

'Access to Information', *Centre for Environment, Fisheries & Aquaculture Science (CEFAS): the Department for Environment, Food and Rural Affairs (DEFRA) Website* <http://www.cefas.defra.gov.uk/publications-and-data/access-to-information.aspx> [correct on and last accessed 1 November 2014]

'ADS Terms and Conditions (September 2014)', *The Archaeology Data Service (ADS) Website* <http://archaeologydataservice.ac.uk/advice/termsOfUseAndAccess> [accessed 9 August 2015]

‘African Oral History: Oral history across generations – A research programme with the universities of Dakar and Algiers’. *University of Portsmouth Website* <http://www.port.ac.uk/research/africanoralhistory/> [accessed 9 August 2015]

‘Announcement: Launch of an online data journal’, *Nature*, 502, (142) (10 October 2013) <http://dx.doi.org/10.1038/502142a>

‘Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0)’, *Creative Commons Website* <http://creativecommons.org/licenses/by-nc-sa/3.0/> [accessed 9 August 2015]

‘Berlin 3 Open Access: Progress in Implementing the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities’, 28 February-1 March 2005, University of Southampton <http://www.eprints.org/events/berlin3/outcomes.html> [accessed 9 August 2015]

‘Berne Convention: For the Protection of Literary and Artistic Works’, *World Intellection Property Organization (WIPO) Website* <http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html> [accessed 9 August 2015]

‘Bethesda Statement on Open Access Publishing’, *Bethesda Open Access Website* <http://www.earlham.edu/~peters/fos/bethesda.htm#summary> [accessed 9 August 2015]

‘Better access to scientific articles on EU-funded research: online pilot project’, *European Union Focus,* 240 (2008), 24-25

‘Centre for Life History and Life Writing Research’, *University of Sussex Website* <http://www.sussex.ac.uk/clhlwr/> [accessed 9 August 2015]

‘CIF – Crystallographic Information Framework’, *Digital Curation Centre (DCC) Website* <http://www.dcc.ac.uk/resources/metadata-standards/cif-crystallographic-information-framework> [accessed 9 August 2015]

‘CIF’, *International Union of Crystallography Website* <http://www.iucr.org/resources/cif> [accessed 9 August 2015]

‘Code of Conduct and Best Practice Guidelines for Journal Editors’, (2011) *Committee on Publication Ethics (COPE) Website* <http://publicationethics.org/files/Code_of_conduct_for_journal_editors.pdf> [accessed 9 August 2015]

‘Crystallography Collection’, *Royal Institute Channel Website* <http://www.richannel.org/collections/2013/crystallography> [accessed 9 August 2015]

‘Crystallography timeline: Explore the history of one of the greatest innovations of the twentieth century’. *Royal Institute Channel Website* <http://www.rigb.org/our-history/history-of-research/crystallography-timeline> [accessed 9 August 2015]

‘Data Catalog Vocabulary’ (DCAT), *W3C Website* <http://www.w3.org/TR/vocab-dcat/> [accessed 9 August 2015]

‘Data Policy’, *NERC Website* <http://www.nerc.ac.uk/research/sites/data/policy/> [accessed 9 August 2015]

‘Database Contents License (DbCL) v1.0’, *Open Data Commons Website* <http://opendatacommons.org/licenses/dbcl/1.0/> [accessed 9 August 2015]

‘Department of Language and Linguistics: Academic Staff – Professor Florence Myles’, *University of Essex Website* <http://www.essex.ac.uk/langling/staff/profile.aspx?ID=2332> [accessed 9 August 2015]

‘Disgraced S Korean cloner Hwang back with Coyote Claim’, *BBC News*, 17 October 2011 <http://www.bbc.co.uk/news/world-asia-pacific-15340240> [accessed 9 August 2015]

‘Mammoth Task: Plan to Clone Ice Age Beast’, *Sky News*, 13 March 2012 <http://news.sky.com/story/2931/mammoth-task-plan-to-clone-ice-age-beast> [accessed 9 August 2015]

‘DOI Handbook: 8 Registration Agencies’, *International DOI Foundation Website* <http://www.doi.org/doi_handbook/8_Registration_Agencies.html> [accessed 9 August 2015]

‘Edinburgh DataShare’, *University of Edinburgh Website* <http://www.ed.ac.uk/schools-departments/information-services/research-support/data-library/data-repository> [accessed 9 August 2015]

‘Editorial: Standards for papers on cloning’, *Nature*, 432 (243) (19 January 2006) <http://dx.doi.org/10.1038/439243a>

‘Exceptions to Copyright: Research’, UK Intellectual Property Office Document (October 2014), *UK Government Website* <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/315014/copyright-guidance-research.pdf> [accessed 9 August 2015]

‘GEMINI’, *Association for Geographical Information (AGI) Website* <http://www.agi.org.uk/uk-gemini/> [accessed 9 August 2015]

‘Guidance: Lambert Toolkit’, *UK Intellectual Property Office: UK Government Website* <https://www.gov.uk/lambert-toolkit> [accessed 9 August 2015]

‘Health and support’, *University of Southampton Website* <http://www.southampton.ac.uk/undergraduate/studentlife/healthandsupport.html> [accessed 9 August 2015]

‘Information about CIF Format Required’, *The Royal Society of Chemistry Website* <http://www.rsc.org/Publishing/Journals/guidelines/AuthorGuidelines/Authoring

Tools/CIFDataImporter/CIFFormatForCIfDataImporter.asp> [accessed 9 August 2015]

'ISO 19115-1:2014 Geographic Information – Metadata—Part 1: Fundamentals', *International Organization for Standardization (ISO) Website* <http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798> [accessed 9 August 2015]

'Jmol: an open-source Java viewer for chemical structures in 3D', *Jmol Website* <http://jmol.sourceforge.net/> [accessed 9 August 2015]

'Legal', *The Met Office Website* <http://www.metoffice.gov.uk/about-us/legal> [accessed 9 August 2015]

'Mark-up validation service', *W3C Website* <http://validator.w3.org/check?uri=http%3a%2f%2fwww%2eflloc%2esoton%2eac%2euk%2f> [accessed 9 August 2015]

'Memories of the Cuban Revolution', *University of Southampton Website* <http://www.southampton.ac.uk/cuban-oral-history/> [accessed 9 August 2015]

'More about interdisciplinarity', *University College London Website* <http://www.ucl.ac.uk/basc/faq/interdisciplinarity> [accessed 9 August 2015]

'National Life Stories', *British Library Website* <http://www.bl.uk/nls> [accessed 9 August 2015]

'NERC Policy on Licensing and Charging for Environmental Data and Information Products', *Natural Environmental Research Council (NERC) Website* <http://www.nerc.ac.uk/research/sites/data/policy/nerc-licensing-charging-policy.pdf> [accessed 9 August 2015]

'Nurses' lives: the oral history of nurses', *Kingston University, Faculty of Health and Social Care Website* <http://www.healthcare.ac.uk/research/nurses-lives/> [accessed 9 August 2015]

'Open Access and ePrints Soton: Introduction', *University of Southampton Website* <http://library.soton.ac.uk/openaccess> [accessed 9 August 2015]

'Open Access to Knowledge in the Sciences and Humanities', Location: Harnack House of the Max Planck Society, Berlin-Dahlem, Germany (20-22 October 2003), *Open Access Max-Planck-Gesellschaft Website* <http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/> [accessed 9 August 2015]

'Open Data from UK Academic Institutions <http://hub.data.ac.uk/> [accessed 9 August 2015]

'Open Government Licence for public sector information 3.0', *The National Archives Website* <https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/> [accessed 9 August 2015]

‘Open Malaria Project’, *Open Source Malaria Website* <http://opensourcemalaria.org/> [accessed 9 August 2015]

‘Open Research Data Handbook – Call for case Studies’, 9 April 2013, *The Open Knowledge Foundation Blog Website* <http://blog.okfn.org/2013/04/09/open-research-data-handbook-call-for-case-studies/#sthash.2r4RbzuY.dpuf> [accessed 9 August 2015]

‘OS OpenData’, *Ordnance Survey (OS) Website* <http://www.ordnancesurvey.co.uk/business-and-government/products/opendata-products.html> [accessed 9 August 2015]

‘Pilot portal for bathymetry’, *European Marine Observation and Data Network (EMODnet) Website* <http://www.emodnet-hydrography.eu/> [accessed 9 August 2015]

‘PLOS Editorial and Publishing Policies’, *PLOS ONE Website* <https://www.plos.org/policies/> [accessed 9 August 2015]

‘PROV-DM: The PROV Data Model’, *W3C Website* <http://www.w3.org/TR/prov-dm/> [accessed 9 August 2015]

‘RCUK announces block grants for universities to aid drives to open access to research outputs’, *Research Councils UK’ (RCUK) Website,* 8 November 2012 <http://www.rcuk.ac.uk/media/news/121108/> [accessed 9 August 2015]

‘Research and Enterprise Services: Ethics in the University’, *The Newcastle University Website* <http://www.ncl.ac.uk/res/research/ethics_governance/ethics/procedures/university/ethics_university.htm> [accessed 9 August 2015]

‘Science as an open enterprise: case studies’, *The Royal Society Website* <http://royalsociety.org/policy/projects/science-public-enterprise/case-studies/> [accessed 9 August 2015]

‘Science as an open enterprise’, The Royal Society Science Policy Centre Report, 02/12 (June 2012), *The Royal Society Website* <http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf> [accessed 9 August 2015]

‘Scottish Oral History Centre’, *University of Strathclyde Website* <http://www.strath.ac.uk/humanities/research/history/sohc/> [accessed 9 August 2015]

‘Sound collection’, *The Imperial War Museum Website* <http://www.iwm.org.uk/collections-research/about/sound> [accessed 9 August 2015]

‘South Korean and Russian scientists bid to clone mammoth’, *Telegraph,* 13 March 2012 <http://www.telegraph.co.uk/earth/wildlife/9139976/South-Korean-and-Russian-scientists-bid-to-clone-mammoth.html> [accessed 9 August 2015]

‘STAP retracted: Two retractions highlight long-standing issues of trust and sloppiness that must be addressed’, *Nature,* 511 (7507), (2 July 2014), 5-6 <http://dx.doi.org/10.1038/511005b>

‘Structure-Property Mapping: Combination Chemistry & the grid (CombiChem)’, Research grant, EPSRC Reference: GR/R67729/0, *Engineering and Physical Sciences Research Council (EPSRC) Website* <http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=GR/R67729/01> [accessed 9 August 2015]

‘Supporting Online Material for Responding to Fraud’, *Science,* 314 (5804) (1 December 2006) <http://dx.doi.org/10.1126/science.1137840>

‘The birth of the Web, *European Organisation for Nuclear Research (CERN) Website* <http://home.web.cern.ch/topics/birth-web> [accessed 9 August 2015]

‘The Language Archive: Team – Expertise’, *Max Planck Institute for Psycholinguistics Website* <http://tla.mpi.nl/team/expertise/> [accessed 9 August 2015]

‘The Southampton Diffraction Centre’, *University of Southampton Website* <http://www.southampton.ac.uk/sdc/> [accessed 9 August 2015]

‘The Statute of Anne; April 10, 1710’, *The Avalon Project Lillian Goldman Law Library, Yale Law School, Yale University Website* <http://avalon.law.yale.edu/18th_century/anne_1710.asp> [accessed 9 August 2015]

‘Tim Berners-Lee’s Proposal’, *European Organisation for Nuclear Research (CERN) Website* <http://info.cern.ch/Proposal.html> [accessed 9 August 2015]

‘Timeline’, *Open Access Directory Website* <http://oad.simmons.edu/oadwiki/Timeline> [accessed 9 August 2015]

‘UK Hydrographic Office INSPIRE Portal and MEDIN Bathymetry Data Archive Centre’, *United Kingdom Hydrographic Office (UKHO) Website* <http://www.ukho.gov.uk/inspire/Pages/home.aspx> [accessed 9 August 2015]

‘University of Southampton Open Data Service’, *University of Southampton Website* <http://data.southampton.ac.uk/> [accessed 9 August 2015]

‘Web Science: how the Web is changing the world’, *Future Learn Website* <https://www.futurelearn.com/courses/web-science> [accessed 9 August 2015]

‘Web Science’, *University of Southampton Website* <http://www.southampton.ac.uk/webscience> [accessed 9 August 2015]

Aliprandi, Simone. 2012. 'Open licensing and databases', *International Free and Open Source Software Law Review*, 4 (1), 5-18 <http://www.ifosslr.org/ifosslr/article/view/62/116> [accessed 9 August 2015]

Amaratunga, Dilanthi and David Baldry. 2001. 'Case study methodology as a means of theory building: performance measurement in facilities management organisations', *Work Study*, 50 (3), 95-105 <http://dx.doi.org/10.1108/00438020110389227>

Angelopoulos, Christina J. 2008. 'Modern intellectual property legislation: warm for reform', *Entertainment Law Review*, 19 (2), 35-40

Badia, Antonio. 2012. 'Evaluating Source Trustability with Data Provenance: A Research Note', *Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI)*, Date: 11-14 June, Location: Washington D.C., USA, pp. 129-131 <http://dx.doi.org/10.1109/ISI.2012.6284145>

Baehr, Craig and Bob Schaller. 2010. *Writing for the Internet: A Guide to Real Communication in Virtual Space* (California: Greenwood Press). Google eBook

Ball, Alex. 2011. 'Smart Research Framework', 1 April, *UKOLN Informatic Group, University of Bath Website* <http://irg.ukoln.ac.uk/2011/04/01/smart-research-framework/> [accessed 9 August 2015]

Bartlett, Oliver. 2013. 'Linked Data: Connecting together the BBC's Online Content', *BBC Blog*, 19 February <http://www.bbc.co.uk/blogs/internet/posts/Linked-Data-Connecting-together-the-BBCs-Online-Content> [accessed 9 August 2015]

Baty, Phil. 2012. 'Leader: Box ticked, but job not yet done', *The Times Higher Education*, 12 July <http://www.timeshighereducation.co.uk/story.asp?storycode=420544> [accessed 9 August 2015]

Bearman, David A. and Richard H. Lytle. 1985-1986. 'The Power of the Principle of Provenance', *Archiveria*, 21 (Winter), 14-27 <http://journals.sfu.ca/archivar/index.php/archivaria/article/view/11231/12170> [accessed 9 August 2015]

Bebbington, Laurence W. 2001. 'Managing content: licensing, copyright and privacy issues in managing electronic resources', *Legal Information Management*, 1 (2), 4-13

Bellinger, Gene, Durval Castro and Anthony Mills. 2004. 'Data, Information, Knowledge, and Wisdom' *Systems Thinking Online Article, Mental Model Musings Website* <http://www.systems-thinking.org/dikw/dikw.htm> [accessed 9 August 2015]

Benkler, Yochai. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedoms* (London: Yale University Press)

Bent Flyvbjerg. 2006. 'Five Misunderstandings About Case-Study Research', *Qualitative Inquiry*, 12, 219-245 <http://dx.doi.org/10.1177/1077800405284363>

Bently, Lionel. and Martin Kretschmer (eds.) *Primary Sources on Copyright (1450-1900) Website* <www.copyrighthistory.org> [accessed 9 August 2015]

Bernard, Harvey Russell and Gery W. Ryan. 2010. *Analyzing Qualitative Data: System: Systematic Processes* (London: SAGE Publications). Google eBook

Berners-Lee, Tim and Nigel Shadbolt. 2010. 'Our manifesto for government data', *Guardian*, 21 January <http://www.theguardian.com/news/datablog/2010/jan/21/timbernerslee-government-data> [accessed 9 August 2015]

—— 2011. 'There's gold to be mined from all our data'. *The Times*, 31 December <http://www.thetimes.co.uk/tto/opinion/columnists/article3272618.ece> [accessed 9 August 2015]

Berners-Lee, Tim. 1989. 'Information Management: A Proposal', *Original Proposal for a Global Hypertext Project at CERN, W3 Website* <http://www.w3.org/History/1989/proposal.html> [accessed 9 August 2015]

—— 1997. 'The "Oh yeah?" button'. W3C: Design Issues, Status: personal view, 6 February, *W3C Website* <http://www.w3.org/DesignIssues/UI.html#OhYeah> [accessed 9 August 2015]

—— 2009. 'Linked data'. W3C: Design Issues, Status: personal view, 18 June, *3C Website* <http://www.w3.org/DesignIssues/LinkedData.html> [accessed 9 August 2015]

Berners-Lee, Tim. 2001. James Hendler and Ora Lassila, 'The Semantic Web', *Scientific American*, May, 34-43 <http://www.scientificamerican.com/article/the-semantic-web/> [accessed 9 August 2015]

Berners-Lee, Tim., Wendy Hall and Nigel Shadbolt. 2006. 'The Semantic Web Revisited', *IEEE Intelligent Systems*, May/June, 96-101 <http://eprints.ecs.soton.ac.uk/12614/1/Semantic_Web_Revisted.pdf> [accessed 9 August 2015]. ePrint.

Bhattacharjee, Yudhijit (ed.) 2006. 'Newsmakers: Movers – A Second Chance', *Science*, 313 (1 September), 1233

*Biodiversity Data Journa*l, *Biodiversity Data Journal Website* <http://biodiversitydatajournal.com/> [accessed 9 August 2015]

Birrell, Augustine. 1899. *Seven Lectures on the Law and History of Copyright in Books* (London: Cassell and Company) <http://archive.org/stream/cu31924029522061#page/n5/mode/2up> [accessed 9 August 2015]

Bizer, Christian and others. 2008. 'Linked Data on the Web', *Proceeding of the 17th international conference on World Wide Web*, Session: Workshops, Date: 21-25 April, Location: Beijing, China, 1265-1266 <http://doi.acm.org/10.1145/1367497.1367760>

*BMC Research Notes* (publishes datasets), *BioMed Central Website* <http://www.biomedcentral.com/bmcresnotes/> [accessed 9 August 2015]

Bohannon, John. 2013. 'Who's Afraid of Peer Review?' *Science*, 342 (6154) (4 October), 60-65 <http://dx.doi.org/10.1126/science.342.6154.60>

*British Association for Applied Linguistics (BAAL) Website* <http://www.baal.org.uk/> [accessed 9 August 2015]

Brown, Mark. 2011. 'Open Access at the University of Southampton: Pushing the boundaries and the art of the possible – Case study', *JISC case study report*, Doc# 796, Version 1.1 (October) <http://www.jisc.ac.uk/media/documents/topics/openaccess/JISC_SouthamptonCase_1.pdf> [accessed 9 August 2015]

*Budapest Open Access Initiative Website* <http://www.budapestopenaccessinitiative.org/ > [accessed 9 August 2015]

Burrows, Jeremy. 2013. 'Advancing antimalarial drug research through open source initiatives', *Guardian*, 24 July <http://www.theguardian.com/global-development-professionals-network/2013/jul/24/open-source-drug-discovery-research> [accessed 9 August 2015]

Bush, Vannevar. 1945. 'As we may think'. *Atlantic Monthly*, (1 July) <http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/> [accessed 9 August 2015]

Callaghan, Sarah. 'A list of Data Journals (in no particular order)', *trac Website* <http://proj.badc.rl.ac.uk/preparde/blog/DataJournalsList> [accessed 9 August 2015]

*Caltech THESIS Website* <http://thesis.library.caltech.edu/> [accessed 9 August 2015]

Caroline Howard (ed.) 2005. *Encyclopedia of Distance Learning* (London: Idea Group)

Carson, John M. and Brian C. Leubitz. 2004. 'Case Comment: United States: copyright - protection of databases'. *European Intellectual Property Review*, 26 (5), N74-75

Check, Erica. 2005. ‘Stem-cell scientist asks for retraction: US partner urges Korean cloner to retract landmark paper’, *Nature News*, 14 December <http://dx.doi.org/10.1038/news051212-5>

—— 2006. ‘Journals scolded for slack disclosure rules’, *Nature News*, 18 January <http://dx.doi.org/10.1038/news060116-6>

*Chemical Mark-up Language (CML) Website* <http://www.xml-cml.org/> [accessed 9 August 2015]

*ChemSpider Website* <http://www.chemspider.com/> [accessed 9 August 2015]

Chen, Peter P. and Leah Y. Wong. 2007. *Active Conceptual Modeling of Learning: Next Generation Learning-Base System Development* (Springer Berlin Heidelberg). Springer eBook <http://dx.doi.org/10.1007/978-3-540-77503-4_3>

Cheney, James and others. 2009. ‘Provenance: A Future History’, *The 24th International Conference on Object Oriented Programming, Systems, Languages and Applications (OOPSLA)*, Date: 25-29 October, Location: Florida, USA, pp. 1-8 <http://people.seas.harvard.edu/~chong/pubs/onward09-provenance.pdf> [accessed 9 August 2015]

Chong, Sei and Dennis Normile (with reporting by Gretchen Vogel). 2006. ‘News of the Week: How Young Korean Researchers Helped Unearth a Scandal ...’ *Science*, 311 (5757) (6 January), 22-25 <http://dx.doi.org/10.1126/science.311.5757.22>

Chun Lee, Byeong and others. 2005. ‘Dogs cloned from adult somatic cells’, *Nature*, 436 (4 August), 641 <http://dx.doi.org/ 10.1038/436641a>

Claxton, Larry D. 2005. ‘Scientific authorship: Part 1. A window into scientific fraud?’ *Mutation research/Review in Mutation Research*, 589 (1), 17-30 <http://dx.doi.org/10.1016/j.mrrev.2004.07.003>

*CODATA Data Science Journal, CODATA Website* <http://www.codata.org/publications/data-science-journal> [accessed 9 August 2015]

Colston, Catherine. 2001. ‘*Sui Generis* Database Right: ripe for review?’ *Journal of Information, Law and Technology*, (3) <http://strathprints.strath.ac.uk/629/> [accessed 9 August 2015]. ePrint

*CombeChem Website* <http://www.combechem.org/index.php> [accessed 9 August 2015]

*Committee on Publication Ethics (COPE) Website* < http://publicationethics.org/> [accessed 9 August 2015]

Conole, Gráinne. 2006. 'External evaluation of the eBank project', Final Independent Report (16 December) <http://www.ukoln.ac.uk/projects/ebank-uk/evaluation-report-dec-2006/evaluation-report-december-2006.pdf> [accessed 9 August 2015]

Copyright, Designs and Patents Act 1988

Couzin, Jennifer, Constance Holden and Sei Chong. 2006. 'News of the Week: Hwang Aftereffects Reverberate at Journals', *Science*, 311 (5759) (20 January) 321 <http://dx.doi.org/ 10.1126/science.311.5759.321b>

Crang, Mike. 2002. 'Qualitative methods: the new orthodoxy?' *Progress in Human Geography*, 26 (5), 647-655 <http://dx.doi.org/10.1191/0309132502ph392pr>

*Creative Commons Website* <https://creativecommons.org/> [accessed 9 August 2015]

*CrossRef Website* <http://www.crossref.org/ > [accessed 9 August 2015]

*CrystalEye Website* <http://wwmm.ch.cam.ac.uk/crystaleye/> [correct on and last accessed 1 December 2013]

*Crystallography Open Database Website* <http://www.crystallography.net/> [accessed 9 August 2015]

Cyganiak, Richard and Anja Jentzsch, *Linked Data Cloud Website* <http://lod-cloud.net/> [accessed 9 August 2015]

Cyranoski, David (with additional reporting by Erica Check). 2006. 'Who's who: a quick guide to the people behind the Woo Suk Hwang story', *Nature News*, 11 January <http://dx.doi.org/10.1038/news060109-9>

Cyranoski, David. 2009. 'Woo Suk Hwang convicted, but not of fraud', *Nature News*, 461 (1181) (26 October) <http://dx.doi.org/10.1038/4611181a>

*DataCite Website* <https://www.datacite.org/node> [accessed 9 August 2015]

*Dataset Papers in Science, Hindawi Publishing Corporation Website* <http://www.hindawi.com/journals/dpis/> [accessed 9 August 2015]

De Zwart, Melissa. 2007. 'A historical analysis of the birth of fair dealing and fair use: lessons for the digital age', *Intellectual Property Quarterly*, 1, 60-91

Deazley, Ronan, Martin Kretschmer, and Lionel Bently (eds.) 2010. *Privilege and Property: Essays on the History of Copyright* (Cambridge: Open Book Publishers). Google eBook

Deazley, Ronan. 2004. *On the Origin of the Right to Copy: Charting the Movement of Copyright Law in Eighteenth Century Britain (1695-1775)* (Oxford: Hart Publishing). Google eBook

—— 2006. *Rethinking Copyright: History, Theory, Language* (Cheltenham: Edward Elgar). Google eBook

Denzin, Norman K. and Yvonna S. Lincoln (eds.) 2011. *The SAGE Handbook of Qualitative Research*, 4th edn (London: SAGE Publications). Google eBook

DiCicco-Bloom, Barbara. and Benjamin F. Crabtree. 2006. 'The qualitative research interview', *Medical Education*, 40 (4), 314-321 < http://dx.doi.org/10.1111/j.1365-2929.2006.02418.x>

Dieter Fensel and others (eds.) 2003. *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential* (Massachusetts: MIT Press) <http://www-ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-mit-press-(with-citation).htm> [accessed 9 August 2015]

*Digital Curation Centre (DCC) Website* <http://www.dcc.ac.uk/> [accessed 9 August 2015]

*Doing Phonetics by Computer (Praat) Website* <http://www.fon.hum.uva.nl/praat/> [accessed 9 August 2015]

*Dublin Core Website* <http://dublincore.org/> [accessed 9 August 2015]

Dutfield, Graham M.and Uma Suthersanen. 2004. 'The innovation dilemma: intellectual property and the historical legacy of cumulative creativity', *Intellectual Property Quarterly*, 4, 379-421

Dzeng, Elizabeth. 2013. **'**How to inspire interdisciplinarity: lessons from the collegiate system', *Guardian,* 15 March <http://www.theguardian.com/higher-education-network/blog/2013/mar/15/interdisciplinary-academic-universities-research> [accessed 9 August 2015]

*Earth System Science Data* (first issue 2009), *Earth System Science Data Website* <http://www.earth-system-science-data.net/> [accessed 9 August 2015]

Ebert, Lawrence B. 2006. 'Lessons to be Learned from the Hwang Matter: Analyzing Innovation the Right Way', *Patent and Trademark Office Society*, 88, 239-255

*Ecological Archives, Ecological Society of America Website* <http://esapubs.org/archive/instruct_d.htm> [accessed 9 August 2015]

Eggington, Elaine, Rupert Osborn and Claude Kaplan. 2013. 'Collaborative Research between Businesses and Universities: The Lambert Toolkit 8 Years On', An independent report commissioned by the Intellectual Property Office (IPO) in collaboration with AURIL, CBI, PraxisUnico & TSB and carried out by IP Pragmatics Limited *UK Intellectual Property Office: UK Government Website* <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/311757/ipresearch-lambert.pdf > [accessed 9 August 2015]

Eisenhardt, Kathleen M. 1989. 'Building Theories from Case Study Research', *Academy of Management Review*, 14 (4), 532-550 <http://www.jstor.org/stable/258557> [accessed 9 August 2015]

Eisenstein, Elizabeth L. 2005. *The Printing Revolution in early Modern Europe* 2nd edn (New York, NY: Cambridge University Press). Google eBook

Elwood, Sarah A. and Deborah G. Martin. 2000. '"Placing" interviews: location and scales of power in qualitative research', *Professional Geographer,* 52 (4), 649-57 <http://www.utsc.utoronto.ca/~kmacd/IDSC10/Readings/interviews/place.pdf> [accessed 9 August 2015]

*ePrints Soton Website* <http://eprints.soton.ac.uk/> [accessed 9 August 2015]

*EPrints Website* <http://www.eprints.org/> [accessed 9 August 2015]

*Erasmus+ Website* <https://erasmusplus.org.uk/llp-and-youth-in-action/erasmus> [accessed 9 August 2015]

*Ethics and Research Governance Online (ERGO), University of Southampton Website* <https://www.ergo.soton.ac.uk/> [accessed 9 August 2015]

Evans, G.R. 2008. 'Academic Libraries and the Law: What Legal Protections Guarantee the Survival of Britain's Academic Library Collections?' *Education Law,* 4, 248

*F1000 Research, F1000 Research Website* <http://f1000research.com/> [accessed 4 July 2015]

Fanelli, Daniele. 2013. 'Redefine misconduct as distorted reporting', *Nature*, 494 (149) (14 February) <http://dx.doi.org/10.1038/494149a>

—— 2013. 'Why Growing Retractions Are (Mostly) a Good Sign', *PLoS Medicine*, 10 (12) (2013) e1001563 <http://dx.doi.org/10.1371/journal.pmed.1001563>

Farchy, Joelle. 2009. 'Are free licences suitable for cultural works?' *European Intellectual Property Review*, 31 (5), 255-263

*figshare Website* <http://figshare.com/ > [accessed 9 August 2015]

Fox, Mark, Tony Ciro and Nancy Duncan. 2005. 'Creative Commons: an alternative, web-based copyright system', *Entertainment Law Review*, 16 (5), 111-116

Fuyuno, Ichiko. 2006. 'Business: Hwang scandal hits Korean biotech hard', *Nature,* 439 (19 January), 265 <http://dx.doi.org/10.1038/439265a>

Gallagher, James. 2014. 'Japanese stem-cell 'breakthrough' findings retracted', *BBC News*, 2 July <http://www.bbc.co.uk/news/health-28124749> [accessed 9 August 2015]

Ganea, Peter, Thomas Pattloch and Christopher Heath (eds.) 2005. *Intellectual Property Law in China* (The Hague, the Netherlands: Kluwer Law International). Google eBook

Gee, David. 2008. 'Should librarians and information professionals be content with current UK copyright law?' *Legal Information Management,* 8 (3), 204-213

*Geoscience Data Journal* (first issue published in June 2014), *Wiley Website* <http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060> [accessed 9 August 2015]

Gibbs, Graham. 2007. *Analyzing Qualitative Data* (London: SAGE Publications). Google eBook

Gibson, C. 1998. 'Semi-structured and unstructured interviewing: a comparison of methodologies in research with patients following discharge from an acute psychiatric hospital', *Journal of Psychiatric and Mental Health Nursing,* 5 (6), 469-477 <http://dx.doi.org/10.1046/j.1365-2850.1998.560469.x>

*GigaScience Journal*, *GigaScience Journal Website* <http://www.gigasciencejournal.com/> [accessed 9 August 2015]

Gilbert, Natasha. 2009. 'Editor will quit over hoax paper: Computer-generated manuscript accepted for publication in open-access journal', *Nature News*, 15 June <http://dx.doi.org/10.1038/news.2009.571>

Gill, John. 2012. 'Times Higher Education Awards 2012 winners', *The Times* <http://www.timeshighereducation.co.uk/story.aspx?storyCode=2003147> [accessed 9 August 2015]

Gillespie, Alexandra. 2006. *Print Culture and the Medieval Author: Chaucer, Lydgate, and their Books 1473-1557* (Oxford: Oxford University Press)

Glaser, Barney G. and Anselm L. Strauss. 2012. *The Discovery of Grounded Theory: Strategies for Qualitative Research,* 7th paperback printing (New Jersey, USA: Transaction Publishers). Google eBook.

Gompel, Stef Van. 2007. 'Unlocking the potential of pre-existing content: how to address the issue of orphan works in Europe?' *International Review of Intellectual Property and Competition Law*, 38 (6), 669-702

Gorman, G.E. 2005. *International Yearbook of Library and Information Management 2004-2005: Scholarly Publishing in an Electronic Era* (London: Facet Publishing)

Greenland, Philip and Phil B. Fontanarosa. 2012. 'Editorial: Ending Honorary Authorship', *Science* 337 (6098) (31 August), 1019 <http://dx.doi.org/10.1126/science.1224988>

Gross, Katherine L. and Gary G. Mittelbach. 2008-9. 'What Maintains the Integrity of Science: An Essay for Nonscientists', *Emory Law Journal*, 58, 341-356

Günergun, Feza and Dhruv Raina (eds.) 2011. *Science between Europe and Asia: Historical studies on the transmission, adoption and adaption of knowledge* (London: Springer Dordrecht Heidelberg). Google eBook

Halfpenny, Peter. 1979. 'The analysis of qualitative data', *The Sociological Review*, 27 (4), 799-827 <http://dx.doi.org/10.1111/j.1467-954X.1979.tb00361.x>

Hall, Sydney R., Frank H. Allen, and I. David Brown. 1991. 'The Crystallographic Information File (CIF): A New Standard Archive File for Crystallography', *Acta Crystallographica*, A47, 655-685 <http://www.iucr.org/__data/iucr/cif/standard/cifstd1.html> [accessed 9 August 2015]

Hamilton, Marci A. 1996. 'The TRIPS Agreement: Imperialistic, Outdated and Overprotective', *Vanderbilt Journal of Transnational Law*, 29, 616-617

Hardy, Melissa and Alan Bryman. 2004. *Handbook of Data Analysis* (London: SAGE Publications). Google eBook

Harnad, Steven and others. 2004. 'The Access/Impact Problem and the Green and Gold Roads to Open Access', *Serials Review*, 30 (4) <http://dx.doi.org/10.1016/j.serrev.2004.09.013>

Hendler, James and others. 2008. 'Web science: an interdisciplinary approach to understanding the web', *Communications of the ACM*, 51 (7), 60-69 <http://dx.doi.org/10.1145/1364782.1364798>

Hendler, James. 2012. 'Developers: A Primer on Machine Readability for Online Documents and Data', 24 September, *USA Government Data Website: Developers* <https://www.data.gov/developers/blog/primer-machine-readability-online-documents-and-data> [accessed 9 August 2015]

Hilbert, Martin and Priscila López. 2011. 'The World's Technological Capacity to Store, Communicate, and Compute Information', *Science*, 322 (6025), 60-65 <http://dx.doi.org/10.1126/science.1200970>

Hitchcock, Steve. 2012. 'Trialling DataCite for chemistry lab notebooks and repository data services', 18 December, *Jisc DataPool Project Website* <http://datapool.soton.ac.uk/2012/12/18/trialling-datacite-for-chemistry-lab-notebooks-and-repository-data-services/> [accessed 9 August 2015]

Holden, Constance (with reporting by Gretchen Vogel and Dennis Normile). 2005. 'News: Korean Cloner Admits Lying About Oocyte Donations', *Science*, 310 (5753) (2 December), 1402-1403 <http://dx.doi.org/10.1126/science.310.5753.1402>

Holden, Constance. 2007. 'News of the Week: Former Hwang Colleague Faked Monkey Data, U.S. Says', *Science*, 315 (5810) (19 January), 317 <http://dx.doi.org/10.1126/science.315.5810.317a>

Horton, Laurence. 2015. 'Digital Object Identifiers: Stability for citations and referencing, but not proxies for quality', *The Impact Blog: The London School of Economics and Political Science*, 23 April <http://blogs.lse.ac.uk/impactofsocialsciences/2015/04/23/digital-object-identifiers-stability-for-citations/> [accessed 9 August 2015]

Hughes, Dave. 2014. 'The new research exemption - where information law and IP collide', *Freedom of Information*, 10 (6), 5-7

Hwang, Woo Suk and others. 2004. 'Evidence of a Pluripotent Human Embryonic Stem Cell Line Derived from a Cloned Blastocyst', *Science*, 303 (5664), 1669-1674 <http://dx.doi.org/10.1126/science.1094515>

—— 2005. 'Patient-Specific Embryonic Stem Cells Derived from Human SCNT Blastocysts', *Science*, 308 (5729), 1777-1783 <http://dx.doi.org/10.1126/science.1112286>

*Infrastructure for Spatial Information in Europe (INSPIRE) Directive Website* <http://inspire.ec.europa.eu/> [accessed 9 August 2015]

Intellectual Property Act 2014

*International DOI Foundation Website* <http://www.doi.org/> [accessed 9 August 2015]

*International Journal of Robotics* (publishes data papers), *SAGE Journals Website* <http://ijr.sagepub.com/> [accessed 9 August 2015]

Jackson, Matt. 2002. 'From Private to Public: Reexamining the Technological Basis for Copyright', *Journal of Communication*, 53 (2), 416-433 <http://dx.doi.org/10.1111/j.1460-2466.2002.tb02553.x>

Jhalani, Mukta. 2008. 'Protecting Egg Donors and Human Embryos – The Failure of the South Korean Bioethics and Biosafety Act', *Pacific Rim Law and Policy Journal*, 17 (3), 707-733

*Jisc DataPool Project Website* <http://datapool.soton.ac.uk/about/> [accessed 9 August 2015]

Johns, Adrian. 1998. *The Nature of the Book: Print and Knowledge in the Making* (Chicago: The University of Chicago Press). Google eBook

Jolley, Jeremy. 2013. *Introducing Research and Evidence-Based Practice for Nursing and Healthcare Professionals*, 2nd edn (Oxford: Routledge). Google eBook

*Journal of Chemical and Engineering Data, ACS Publications Website* <http://pubs.acs.org/journal/jceaax> [accessed 9 August 2015]

*Journal of Open Archaeology Data, Journal of Open Archaeology Data Website* <http://openarchaeologydata.metajnl.com/> [accessed 9 August 2015]

*Journal of Open Psychology Data, Journal of Open Psychology Data Website* <http://openpsychologydata.metajnl.com/> [accessed 9 August 2015]

*Journal of Open Public Health Data, Journal of Open Public Health Data Website* <http://openhealthdata.metajnl.com/> [accessed 9 August 2015]

*Journal of Physical and Chemical Reference Data, AIP Scitation Website* <http://scitation.aip.org/content/aip/journal/jpcrd/browse> [accessed 9 August 2015]

Jump, Paul. 2014. 'Former member's misconduct causes third retraction for lab', *The Times Higher Education*, 13 February <http://www.timeshighereducation.co.uk/news/former-members-misconduct-causes-third-retraction-for-lab/2011262.article> [accessed 9 August 2015]

Kansa, Eric C., Jason Schultz and Ahrash N. Bissell. 2005. 'Protecting traditional knowledge and expanding access to scientific data: juxtaposing intellectual property agendas via a "some rights reserved" model'. *International Journal of Cultural Property*, 12 (3), 285-314

Kapitzke, Cushla. 2001. 'Ceremony and cybrary: Digital libraries and the dialectic of place and space', *Social Alternatives*, 20 (1), 33-40 <http://eprints.qut.edu.au/43995/1/Kapitzke_cybrary_Social_Alternatives.pdf> [accessed 9 August 2015].ePrint

Keen, Paul. 1999. *The crisis of literature in the 1790s: print culture and the public sphere* (Cambridge: Cambridge University Press)

Kennedy, Donald. 2006. 'Editorial Retraction: Retraction of Hwang *et al.*, *Science* 308 (5729) 1777-1783', *Science*, 311 (5759) (12 January), 335 <http://dx.doi.org/10.1126/science.1124926>

—— 2006. 'Editorial: Responding to Fraud', *Science*, 314 (5804) (1 December) 1353 <http://dx.doi.org/10.1126/science.1137840>

Kishani Mendis, Dinusha. 2009. *Universities and Copyright Collecting Societies* (Cambridge: Cambridge University Press)

Kluger, Jeffrey. 2004. 'Scientists and thinkers: the Korean cloners', *TIME*, 26 April <http://www.time.com/time/specials/packages/article/0,28804,1970858_1970909_1971678,00.html> [accessed 9 August 2015]

*Knowledge Exchange Website* <http://www.knowledge-exchange.info/> [accessed 9 August 2015]

*Ladbroke (Football) Ltd v. William Hill (Football) Ltd* [1964] 1 WLR 273

Lea, Gary. 1996. 'In defence of originality', *Entertainment Law Review*, 7 (1), 21-26

León, Alexander and others. 2010. 'Geographical Linked Data: a Spanish Use Case', *Proceedings of the 6th International Conference on Semantic Systems, Session: Pragmatic Web, Conference track: triplification challenge*, Date: 1–3 September, Location: Messe Congress Graz, Austria, 36, 1-3 <http://doi.acm.org/10.1145/1839707.1839753>

Loewenstein, Joseph. 2002. *The Author's Due: Printing and the Prehistory of Copyright* (Chicago: The University of Chicago Press). Google eBook

Lynch, Clifford A. 2001. 'When Documents Deceive: Trust and Provenance as New Factors for Information Retrieval in a Tangled Web', *Journal of the American Society for Information Science and Technology*, 52 (1), 12-17 <http://dx.doi.org/10.1002/1532-2890(2000)52:1<12::AID-ASI1062>3.0.CO;2-V>

Lyon, Liz and others. 2008. 'Scaling Up: Towards a Federation of Crystallography Data Repositories', UK eBank Report, Version 1.0: Final (12 May) <http://www.ukoln.ac.uk/projects/ebank-uk/dissemination/Ebank3report/Ebank3report.pdf> [accessed 9 August 2015]

Lyon, Liz. 2003. 'eBank UK: Building the Links Between Research Data, Scholarly Communication and Learning', *Ariadne*, 36 <http://www.ariadne.ac.uk/issue36/lyon/> [accessed 9 August 2015]

MacWhinney, Brian. 2015. 'The CHILDES Project: Tools for Analyzing Talk – Part 1: The CHAT Transcription Format', *CHILDES Manual*, Carnegie Mellon University (19 August) <http://childes.talkbank.org/manuals/CHAT.pdf> [accessed 9 August 2015]

—— 2015. 'The CHILDES Project: Tools for Analyzing Talk – Part 2: The CLAN programs', *CHILDES Manual*, Carnegie Mellon University (23 June) <http://childes.psy.cmu.edu/manuals/CLAN.pdf> [accessed 9 August 2015]

Marcus, Adam and Ivan Oransky. 'Retraction Watch Blog', *Retraction Watch Website* <http://retractionwatch.wordpress.com/> [accessed 9 August 2015]

—— 2014. 'The Top 10 Retractions of 2014', *The Scientist*, 23 December <http://www.the-scientist.com/?articles.view/articleNo/41777/title/The-Top-10-Retractions-of-2014/> [accessed 9 August 2015]

Marlin-Bennett, Renée. 2004. *Knowledge Power: Intellectual Property, Information and Privacy* (London, Boulder)

Marshall, Catherine and Gretchen B. Rossman. 2006. *Designing Qualitative Research*, 4th edn (London: SAGE Publications)

Marušić, Ana, Lana Bošnjak, and Ana Jerončić. 2011. 'A Systematic Review of Research on the Meaning, Ethics and Practices of Authorship across Scholarly Disciplines', *PLoS One*, 6 (9) (8 September) e23477<http://dx.doi.org/10.1371/journal.pone.0023477>

McCance, Tanya V., Hugh P. McKenna and Jennifer R.P. Boore. 2001. 'Exploring caring using narrative methodology: an analysis of the approach', *Journal of Advanced Nursing*, 33 (3), 350-356 <http://dx.doi.org/10.1046/j.1365-2648.2001.01671.x>

McIlveen, Peter and others. 2003. 'Evaluation of a semi-structured assessment interview derived from systems theory framework', *Australian Journal of Career Development*, 12 (3), 33-41 <http://dx.doi.org/10.1177/103841620301200306>

McIntosh, Carey. 1998. *The Evolution of English Prose, 1700-1800: Style, Politeness, and Print Culture* (Cambridge, Cambridge University Press)

MEDIN Annual Report (2012-13), *'The Marine Environmental Data and Information Network (MEDIN) Website* <http://www.oceannet.org/library/key_documents/documents/medin_annual_report_201213_final.pdf> [accessed 9 August 2015]

Mendelson, Laura L. 2003. 'Privatizing Knowledge: The Demise of Fair Use and the Public University', *Albany Law Journal of Science & Technology*, 13 (2), 593-612

Migheli, Matteo and Giovanni B. Ramello. 2013. 'Open access, social norms and publication choice', *European Journal of Law & Economics*, 35 (2), 149-167

Miles, Matthew B. 1979. 'Qualitative Data as an Attractive Nuisance: The Problem of Analysis', *Administrative Science Quarterly*, 24 (4), 590-601 <http://dx.doi.org/10.2307/2392365>

Moffatt, Suzanne and others. 2006. 'Using quantitative and qualitative data in health services research – what happens when mixed method findings conflict?' *BMC Health Services Research*, 6 (28), 1-10 <http://dx.doi.org/ 10.1186/1472-6963-6-28>

Moreau, Luc and Paul Groth. 2013. *Provenance: An Introduction to PROV* (Morgan & Claypool Publishers). eBook <http://dx.doi.org/10.2200/S00528ED1V01Y201308WBE007>

Moreau, Luc. 2010. 'The Foundations for Provenance on the Web', *Foundations and Trends in Web Science*, 2 (2-3), 99-241 <http://dx.doi.org/10.1561/1800000010> (ePrint version) <http://eprints.soton.ac.uk/271691/> [accessed 9 August 2015]

Morra-Imas, Linda G. and Ray C. Rist. 2009. *The Road to Results: Designing and Conducting Effective Development Evaluations* (Washington DC, USA: The International Bank for Reconstruction and Development/The World Bank). Google eBook

Muniswamy-Reddy, Kiran-Kumar, Peter Macko and Margo Seltzer. 2010. 'Provenance for the Cloud', *FAST'10 Proceedings of the 8th USENIX conference on File and storage technologies* <https://www.usenix.org/legacy/events/fast10/tech/full_papers/muniswamy-reddy.pdf?CFID=450573987&CFTOKEN=30484129> [accessed 9 August 2015]

Murray-Rust, Peter, Cameron Neylon, Rufus Pollock and John Wilbanks. 2010. 'Panton Principles for Open Data in Science', 19 February, *Panton Principles Website* <http://pantonprinciples.org/> [accessed 9 August 2015]

*Nature*, 4 November 1869 <http://www.nature.com/nature/about/first/> [accessed 9 August 2015]

Obokata, Haruko and others. 2014. 'Retraction: Bidirectional developmental potential in reprogrammed cells with acquired pluripotency', *Nature*, 511 (7507), 112 <http://dx.doi.org/10.1038/nature13599>

Ockerbloom, John Mark. 2007. 'Copyright and Provenance: Some Practical Problems', *IEEE Computer Society Technical Committee on Data Engineering*, 30 (4), 51-58 <http://repository.upenn.edu/cgi/viewcontent.cgi?article=1051&context=library_papers> [accessed 9 August 2015]

Omitola, Tope and others. 2011. 'Tracing the Provenance of Linked Data using voiD', *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, Date: 25-27 May, Location: Sogndal, Norway <http://doi.acm.org/10.1145/1988688.1988709>

Opdenakker, Raymond. 2006. 'Advantages and Disadvantages of Four Interview Techniques in Qualitative Research', *Forum: Qualitative Social Research*, 7 (4), 1-13 <http://www.qualitative-research.net/index.php/fqs/article/view/175/392> [accessed 9 August 2015]

*Open Data Commons Website* <http://opendatacommons.org/> [accessed 9 August 2015]

*Open Data Institute (ODI) Website* <http://www.theodi.org/> [accessed 9 August 2015]

*Open Knowledge Foundation Website* <https://okfn.org/> [accessed 9 August 2015]

*Open Malaria Project Website* <http://malaria.ourexperiment.org/> [accessed 9 August 2015]

Oransky, Ivan. 'Embargo Watch Blog', *Embargo Watch Website* <https://embargowatch.wordpress.com/> [accessed 9 August 2015]

*ORCID Website* <http://orcid.org/> [accessed 9 August 2015]

*Oxford Dictionaries Website* <http://www.oxforddictionaries.com/ > [accessed 9 August 2015]

*Padua@research Website* <http://paduaresearch.cab.unipd.it/> [accessed 9 August 2015]

Paul Boersma and David Weenink, 'Praat: doing phonetics by computer [Computer program]'. Version 5.4, *'Doing Phonetics by Computer' (Praat) Website* <http://www.fon.hum.uva.nl/praat/> [accessed 9 August 2015]

Peifer, Karl-Nikolaus. 2008. 'The return of the commons – copyright history as a helpful source?' *International Review of Intellectual Property and Competition Law*, 39 (6), 679-688

*Philosophical Transactions*, 1 (1 January 1665), (1-22) <http://rstl.royalsocietypublishing.org/content/1/1-22.toc> [accessed 9 August 2015]

Pittam, Neil, Stephen Saxby and Chris Hill. 2010. 'Approaches to data policy in the marine sector', *'Marine Environmental Data and Information Network' (MEDIN) Final Project Report,* Version 1.1 (December) <http://www.oceannet.org/library/work_stream_documents/documents/medin_data_policy_study_rep_final_v1_1.pdf> [accessed 9 August 2015]

Pope, Catherine, Sue Ziebland and Nicholas Mays. 2000. 'Analysing qualitative data', *British Medical Journal*, 320 (7227), 114-116 <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1117368/> [accessed 9 August 2015]

Punch, Keith F. 2005. *Introduction to Social Research: Quantitative and Qualitative Approaches*, 2nd edn (London, SAGE Publications). Google eBook.

Quinn Patton, Michael. 2002. *Qualitative Research and Evaluation Methods,* 3rd edn (London: SAGE Publications)

Rabinovitz, Lauren and Abraham Geil (eds.) 2004. *Memory Bytes: History, Technology, and Digital Culture*, (Durham NC: Duke University Press). Google eBook.

Rackham, Les and Rob Walker. 2010. Metadata Guidelines for Geospatial Data Resources - Part 1: Introduction (Association for Geographical Information, September) <http://www.agi.org.uk/uk-gemini/> [accessed 9 August 2015]

Rahn, Kim. 2008. 'Will Hwang Woo-Suk Return? Decision on Permits of Stem Cell Research Due Sat', *The Korean Times Online*, 27 July <http://www.koreatimes.co.kr/www/news/nation/2010/09/117_28274.html> [accessed 9 August 2015]

RCUK, 'RCUK Policy on Open Access and Supporting Guidance', *Research Councils UK (RCUK) Website* <http://www.rcuk.ac.uk/documents/documents/RCUKOpenAccessPolicy.pdf> [accessed 9 August 2015]

*Research Data Alliance Website* <https://rd-alliance.org/> [accessed 9 August 2015]

*Research Governance Office (RGO), University of Southampton Website* <http://www.southampton.ac.uk/corporateservices/rgo/> [accessed 9 August 2015]

Robson, Colin. 2011. *Real World Research*, 3rd edn (Chichester: Wiley)

Rosati, Eleonor. 2013. 'The orphan works provisions of the ERR Act: are they compatible with UK and EU laws?' *European Intellectual Property Review*, 35 (12), 724-740

Rose, Mark. 1993. *Authors and Owners: The Invention of Copyright* (Cambridge MA: Harvard University Press). Google eBook

Ross, Alexander. 2014. 'Copyright works: seeking the lost', *Entertainment Law Review*, 25 (3), 104-107

Sample, Ian. 2014. 'How computer-generated fake papers are flooding academia', *Guardian*, 26 February <http://www.theguardian.com/technology/shortcuts/2014/feb/26/how-computer-generated-fake-papers-flooding-academia> [accessed 9 August 2015]

—— 2014. 'Stem cell scientist Haruko Obokata found guilty of misconduct', *Guardian*, 1 April <http://www.theguardian.com/science/2014/apr/01/stem-cell-scientist-haruko-obokata-guilty-misconduct-committee> [accessed 9 August 2015]

Sang-Hun, Choe. 2006. 'Researcher who faked cloning data gets new job – Asia – Pacific – International Herald Tribune', *The New York Times*, 18 August <http://www.nytimes.com/2006/08/18/world/asia/18iht-clone.2528877.html> [accessed 9 August 2015]

Saxby, Stephen. 1990. *The Age of Information* (London: The Macmillan Press Ltd)

*Scientific Data, Nature Website* <http://www.nature.com/sdata/> [accessed 9 August 2015]

Secker, Jane and Maria Bell. 2010. 'Copyright? Why would I need to worry about that? The challenges of providing copyright support for staff', *Legal Information Management*, 10 (3), 166-170

Seeley, Becky and others. 2014. 'Guidance notes for the production of discovery metadata for the Marine Environmental Data and Information Network (MEDIN)', *MEDIN Report*, version 2.3.8

<http://www.oceannet.org/marine_data_standards/documents/medin_schema_doc_2_3_8_brief.pdf> [accessed 9 August 2015]

Self, Henry. 2002. 'Digital Sampling: A Cultural Perspective', *UCLA Entertainment Law Review*, 9 (2), 347-359

Shepherd, Jessica. 2009. 'Editor quits after journal accepts bogus science article', *Guardian*, 18 June <http://www.guardian.co.uk/education/2009/jun/18/science-editor-resigns-hoax-article> [accessed 9 August 2015]

Sherwood-Edwards, Mark. 1995. 'The redundancy of originality', *Entertainment Law Review*, 6 (3), 94-106

Silver, Ingrid and Helen Anderson. 2010. 'Gutenberg odyssey - the advent of e-books and some implications for the world of publishing', *Entertainment Law Review*, 21 (6), 225-228

Silverman, D. 1998. 'Qualitative research: meanings or practices?' *Information Systems Journal*, 8 (1), 3-20 <http://dx.doi.org/10.1046/j.1365-2575.1998.00002.x>

Simmhan, Yogesh L. and others. 2006. 'Performance Evaluation of the Karma Provenance Framework for Scientific Workflows' in *Provenance and Annotation of Data: International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006, Revised Selected Papers*, ed. by Luc Moreau and Ian Foster (Springer Berlin Heidelberg) 222-236 <http://dx.doi.org/10.1007/11890850_23>

Simmhan, Yogesh L., Beth Plale and Dennis Gannon. 2005. 'A Survey of Data Provenance in e-Science', *SIGMOD Record*, 34 (3), 31-36 <http://www.sigmod.org/publications/sigmod-record/0509/p31-special-sw-section-5.pdf> [accessed 9 August 2015]

Simmonds, Tony. 2010. 'Common knowledge? The rise of Creative Commons licensing', *Legal Information Management*, 10 (3), 162-165

Sokol, Daniel K. 2008. The dilemma of authorship', *British Medical Journal*, 336 <http://dx.doi.org/10.1136/bmj.39500.620174.94>

Spector, Horacio M. 1989. 'An outline of a theory justifying intellectual and industrial property rights', *European Intellectual Property Review*, 11 (8), 270-273

Speight, Dunstan and Jennifer Darroch. 2012. 'Clarifying copyright', *Legal Information Management*, 12 (3), 209-213

Steen, R. Grant, Arturo Casadevall, and Ferric C. Fang. 2013. 'Why Has the Number of Scientific Retractions Increased?' *PLoS One*, 8 (7) <http://dx.doi.org/10.1371/journal.pone.0068397>

Stewart, Jon. 2011. 'Global data storage calculated at 295 exabytes', *BBC*, 11 February <http://www.bbc.co.uk/news/technology-12419672> [accessed 9 August 2015]

Suber, Peter. 2009. 'Timeline of the Open Access Movement', last revised 9 February <http://legacy.earlham.edu/~peters/fos/timeline.htm> [accessed 9 August 2015]

Sullivan, Larry E. 2009. *The Sage Glossary of the Social and Behavioral Sciences* (California: SAGE Publications). Google eBook.

Sun, Haochen. 2005. 'Copyright law under siege: an inquiry into the legitimacy of copyright protection in the context of the global digital divide', *International Review of Intellectual Property and Competition Law*, 36 (2), 192–213

Swan, Alma and Sheridan Brown. 2008. 'To Share or not to Share: Publication and Quality Assurance of Research Data Outputs'. Report undertaken by Key Perspectives Ltd and commissioned by the Research Information Network (RIN). RIN in association with the 'Joint Information Systems Committee' (JISC) and the Natural Environment Research Council (NERC) (June), *The Research Information Network Website* <http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs> [accessed 9 August 2015]

Sweeney, Shelley. 2008. 'The Ambiguous Origins of the Archival Principle of "Provenance"', Libraries *& the Cultural Record*, 43 (2), 193-213 <http://dx.doi.org/10.1353/lac.0.0017>

*TalkBank Website* <http://talkbank.org/> [accessed 9 August 2015]

Tamura, Yoshiyuki. 2009. 'Rethinking copyright institution for the digital age', *WIPO Journal*, 1, 63-74

*The Biological Research Information Center Website* <http://bric.postech.ac.kr/myboard/read.php?id=267&Board=bric_board> [accessed 9 August 2015]

*The British Crystallographic Association Website* <http://crystallography.org.uk/> [accessed 9 August 2015]

*The Cambridge Structural Database, The Cambridge Crystallographic Data Centre, University of Cambridge Website* <http://www.ccdc.cam.ac.uk/Solutions/CSDSystem/Pages/CSD.aspx> [accessed 9 August 2015]

*The Chemical Crystallography Group Website* <http://ccg.crystallography.org.uk/> [accessed 9 August 2015]

The Data Protection Act 1998

*The Economic and Social Research Council (ESRC) Website* <http://www.esrc.ac.uk/> [accessed 9 August 2015]

*The First World Wide Web Conference*, Location: CERN, Geneva, Switzerland (25-27 May 1994) <http://www94.web.cern.ch/WWW94/Welcome.html> [accessed 9 August 2015]

The Open Archives Initiative Protocol for Metadata Harvesting', Protocol Version 2.0 of 2002-06-14. Document Version 2008-12-07T20:42:00Z, *The Open Archives Initiative (OAI) Website* <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm> [accessed 9 August 2015]

*The Smart Tea Project Website* <http://www.smarttea.org/> [accessed 9 August 2015]

*The UK Data Archive Website* <http://www.data-archive.ac.uk/> [accessed 9 August 2015]

*The UK University Council for Modern Languages (UCML) Website* <http://www.ucml.ac.uk/> [accessed 9 August 2015]

*The United Kingdom Office for Library and Information Networking (UKOLN) Website* <http://www.ukoln.ac.uk/> [accessed 9 August 2015]

Thompson Klein, Julie. 2005. *Humanities, Culture, and Interdisciplinarity: The Changing American Academy* (New York, NY: State University of New York Press, Albany) Google eBook

Tonkin, Emma. 2008. 'Persistent Identifiers: Considering the Options', *Ariadne: Web Magazine for Information Professionals*, 30 July <http://www.ariadne.ac.uk/print/issue56/tonkin#15> [accessed 9 August 2015]

*UK Government Legislation Website* <http://www.legislation.gov.uk/ > [accessed 9 August 2015]

*UK Intellectual Property Office, UK Government Website* <https://www.gov.uk/government/organisations/intellectual-property-office/> [accessed 9 August 2015]

*University of Southampton Website* <https://www.southampton.ac.uk/> [accessed 9 August 2015]

*US Department of Commerce: 'National Oceanic and Atmospheric Administration' (NOAA) Website* <http://www.noaa.gov/index.html> [accessed 9 August 2015]

van der Heyden, M. A. G., T. van de ven and T. Opthof. 2009. 'Fraud and misconduct in science: the stem cell seduction', *Netherlands Heart Journal*, 17 (1), 25-29 <http://ukpmc.ac.uk/articles/PMC2626656/> [accessed 9 August 2015]

Van Noorden, Richard. 2014. 'Publishers withdraw more than 120 gibberish papers', *Nature News*, 24 February <http://dx.doi.org/10.1038/nature.2014.14763>

Vogel, Gretchen (with reporting by Dennis Normile). 2006. 'NewsFocus: Picking Up the Pieces After Hwang', *Science*, 312 (5773) (28 April), 516-7 <http://dx.doi.org/10.1126/science.312.5773.516>

Vogel, Gretchen. 2005. 'Landmark Paper Has an Image Problem', *Science*, 310 (5754) (9 December) 1595 <http://dx.doi.org/10.1126/science.310.5754.1595>

'Voices of the University: Memories of Warwick, 1965-2015', *The University of Warwick Website* <http://www2.warwick.ac.uk/fac/cross_fac/ias/current/universityvoices/> [accessed 9 August 2015]

*W3C Website* <http://www.w3.org/People/Berners-Lee/> [accessed 9 August 2015]

Wade, Nicholas. 2006. 'University Panel Faults Cloning Co-Author', *The New York Times*, 11 February <http://www.nytimes.com/2006/02/11/science/11clone.html> [accessed 9 August 2015]

Waelde, Charlotte and Hector MacQueen. 2004. 'From entertainment to education: the scope of copyright?' *Intellectual Property Quarterly*, 3, 259-283

Wager, Elizabeth and Sabine Kleinert. 2012. 'Cooperation between research institutions and journals on research integrity cases', (on behalf of COPE Council) (March), *Committee on Publication Ethics (COPE) Website* <http://publicationethics.org/files/Research_institutions_guidelines_final.pdf> [accessed 9 August 2015]

Warr, Wendy A. 2006. 'Digital Repositories Supporting eResearch: Exploring the eCrystals Federation Model', *Ebank/R4l/Spectra Joint Consultation Workshop*, Location: London, Date: 20 October (December) <http://www.ukoln.ac.uk/projects/ebank-uk/workshops/eBank-SPECTRa-R4L-workshop/eBank-SPECTRa-R4L-workshop.pdf> [accessed 9 August 2015]

*Web Science Trust Website* <http://webscience.org/> [accessed 9 August 2015]

Weber-Wulff, Debora. 'Copy, Shake and Paste: A Blog about Plagiarism and Scientific Misconduct', *Copy, Shake and Paste Website* <http://copy-shake-paste.blogspot.co.uk/> [accessed 9 August 2015]

—— 2012. 'Viewpoint: The spectre of plagiarism haunting Europe', *BBC News*, 25 July <http://www.bbc.co.uk/news/18962349> [accessed 9 August 2015]

Wengraf, Tom. 2001. *Qualitative Research Interviewing* (London: SAGE Publications)

Wohn, D. Yvette and Dennis Normile. 2006. 'Prosecutors Allege Elaborate Deception and Missing Funds', *Science*, 312 (5776) (19 May), 980-981 <http://dx.doi.org/10.1126/science.312.5776.980>

Wood, Larry E. 1997. 'Semi-Structured Interviewing for User-Centered Design', *Interactions*, 4 (2), 48-61 <http://dx.doi.org/10.1145/245129.245134>

Yin, Robert K. 2009. *Case Study Research: Design and Methods*, 4th edn (London: SAGE Publications)

Yu, Peter K. (ed.) 2007. *Intellectual Property and Information Wealth: Issues and Practices in the Digital Age* (Connecticut, USA: Greenwood Publishing Group). Google eBook

Yuan, Jie, Jianya Gong and Mingda Zhang. 2013. 'A Linked Data Approach for Geospatial Data Provenance', *IEEE Transactions on Geoscience and Remote Sensing*, 51 (11), 5105-5112 <http://dx.doi.org/ 10.1109/TGRS.2013.2249523>

Zimmer, Carl. 2012. 'A Sharp Rise in Retractions Prompts Calls for Reform', *New York Times*, 16 April <http://www.nytimes.com/2012/04/17/science/rise-in-scientific-journal-retractions-prompts-calls-for-reform.html?_r=0> [accessed 9 August 2015]