UNIVERSITY OF SOUTHAMPTON

Open Access Analytics with Open Access Repository Data: A Multi-level Perspective

by

Ibraheem M. AL Sadi

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Physical and Applied Sciences
Electronics and Computer Science

January 2021

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL AND APPLIED SCIENCES
ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

OPEN ACCESS ANALYTICS WITH OPEN ACCESS REPOSITORY DATA: A MULTI-LEVEL PERSPECTIVE

by
Ibraheem M. AL Sadi

Within nearly two decades after the open access movement emerged, its community has drawn attention to understanding its development, coverage, obstacles and motivations. To do so, they depend on data-centric analytics of open access publishing activities, using Web information space as their data sources for these analytical activities. Open access repositories are one such data source that nurtures open access publishing activities and are a valuable source for analytics. Therefore, the open access community utilises open access repository infrastructure to develop and operate analytics, harnessing the widely adopted Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) interoperability layer to develop value-added services with an analytics agenda. However, this layer presents its limitations and challenges regarding the support of analytical value-added services.

To address these practices, this research has taken the step to consolidate these practices into the ‘open access analytics’ notion of drawing attention to its significance and bridge it with data analytics literature. As part of this, an explanatory case study demonstrates how the OAI-PMH service provider approach supports open access analytics and also presents its limitations using Registry of Open Access Repositories (ROAR) analytics as a case study. The case study reflects the limitation of open access registries to enable a single point of discovery due to the quality of their records and complexity of open access repositories taxonomy, the complexity of operationalising the unit of analysis in particular analytics due to the limitations in the OAI-PMH metadata schemes, the complex and resource-intensive harvesting process due to the large volume of data and the low quality of OAI-PMH standards adoptions and the issue of service provider suitability due to a single point of failure.

Also, this doctoral thesis proposes the use of Open Access Analytics using Open Access Repository Data with a Social Machine (OAA-OARD-SM) as a conceptual framework to deliver open access analytics by using the open access repository infrastructure in a collaborative manner with social machines. Furthermore, it takes advantage of the web

observatory infrastructure as a form of web-based mediated technology to coordinate the open access analytics process. The conceptual framework re-frames the open access analytics process into four layers: the open access repository layer, the open access registry layer, the data analytics layer and open access analytics layer. It also conceptualises analytics practices carried out within individual repository boundaries as core practices for the realisation of open access analytics and examines how the repository management team can participate in the open access analytics process.

To understand this, expert interviews were carried out to investigate and understand the analytics practices within the repository boundaries and the repository management teams' interactions with analytics applications that are fed by the open access repository or used by repository management to operate open access analytics. The interviews provide insight into the variations in the types of analytic practices and highlight the active role played by the repository management team in these practices. Thus, it provides an understanding of the analytics practices within open access repositories by classifying them into two main categories: the distributed analytical applications and locally operated analytics. The distributed analytics application includes cross-repository OAI-based analytics, cross-repository usage data aggregators, solo-repository content-centric analytics and solo-repository centric analytics. On the other hand, the locally operated analytics take forms of Current Research Information System (CRIS), repository embedded functionalities and in-house developed analytics. It also classifies the repository management interactions with analytics into four roles: data analyst, administrative, data and system management, and system development and support. Lastly, it raises concerns associated with the application of analytics on open access repositories, including data-related, cost-related and analytical concerns.

Contents

List of Figures

List of Tables

Abbreviations

OAR, OARs	Open Access Repository, Open Access Repositories
OARD	Open Access Repository Data
ROAR	Registry of Open Access Repositories
OAA	Open Access Analytics
CORE	COnnecting REpositories
COAR	The Confederation of Open Access Repositories
openDOAR	Directory of Open Access Repositories
WO	Web Observatory
SM	Social Machine
OAA-OARD-SM	Open Access Analytics using pen Access Repository Data with Social Machine
InfoVis	Information Visualisation
VA	Visual ANalytics
KDD	Knowledge Discovery and Data Mining
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting

Declaration of Authorship

I, Ibraheem Al Sadi, 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 on:

Open Access Analytics with Open Access Repository Data: A Multi-level Perspective

I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University;
- 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;
- Where I have consulted the published work of others, this is always clearly attributed;
- 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;
- I have acknowledged all main sources of help;
- 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;
- None of this work has been published before submission;

Signature : ______________________________

Date : ______________________________

Acknowledgements

First and foremost, I would like to thank my great supervisors, Professor Les Carr and Professor Stevan Harnad, to whom I will always be grateful for their incredible guidance, insight, inspiration, and intellectual and emotional support.

I extend my thanks to all the participants who accepted to share with me their interesting experiences, which resulted in the development of this thesis.

My special thanks are due to my employer, the Ministry of Manpower of Oman, for sponsoring my PhD programme. A great deal of thanks goes to the Cultural Attaché of Oman in London, who stood with me during critical stages of my PhD journey.

I owe a deep gratitude to my mother, Dabia Al Sadi, and my father, Mohammad Al Sadi, for all the values and principles she taught me. They have been a source of support, encouragement, and inspiration in every single day of my life.

I extend my thanks to my brothers and sisters, Idress, Mouad, Omar, Habiab, Hassna, Sauad, Fatma, and Alia, for their continuous support and encouragement.

I keep the last sentences for those who share my life, my thoughts, and my feelings. I am extremely thankful to my beloved wife, Iman, my daughter, Shahad, and my son, Mohammad for bearing the hardships of being away from family and friends in Oman. They are my source of happiness and joy. I am especially indebted to my wife for her patience and sacrifices, and bearing the stressful PhD journey. Her continuous support and encouragement will always be held deep in my heart.

To the memory of my FATHER.

He was my inspiration by the smile and the happiness he reflects at any achievements I made.

To the new generation (Shahad & Mohammad).

Chapter 1

Introduction

1.1 Background and Motivations

The dramatic shift in the way people communicate and collaborate due to the existence of the World Wide Web has led many researchers to carry out studies to understand its current state and influence, as well as investigate its growth and dynamic and anticipate its potential. As a result, Berners-Lee et al. (2006) called for the creation of a Web science, describing it as follows:

> Web science is about more than modelling the current Web. It is about engineering new infrastructure protocols and understanding the society that uses them, and it is about the creation of beneficial new systems. It has its own ethos: decentralisation to avoid social and technical bottlenecks, openness to the reuse of information in unexpected ways, and fairness. It uses powerful scientific and mathematical techniques from many disciplines to consider at once microscopic Web properties, macroscopic Web phenomena, and the relationships between them. (Berners-Lee et al., 2006, p 770)

Scholarly communication is one of the sectors that has been highly influenced by the emergence of the Web, resulting in a series of changes to the system as a whole. This began with the digitisation of scholarly work, and electronic updates to the paper-based approach of publishing. What followed were changes in the business model and work-flow of publishing, with preprints, conference papers, data and other materials being archived in *open access repositories* (Awre, 2006). These repositories provide the research community with a long-term preservation mechanism and free access to scientific literature (Lynch, 2003). The sharing and discussion of scholarly works has, therefore, gained a new system of socialisation features such as the scholarly social network and a new generation of repositories. Within this infrastructure, new phenomena arise and are exposed to the Web, providing opportunities for understanding through new tools and methodologies (Van de Sompel and Treloar, 2014).

One of the phenomena that have emerged as a result of the Web is *open access publishing*, which has drawn a considerable amount of attention from the scholarly communication research community (Swan, 2006). The movement, as supported by its open distributed repository infrastructure, pools archiving efforts on a worldwide scale (Awre, 2006). The open access community has accordingly introduced *registries* such as the Registry of Open Access Repositories (ROAR) (Eprints Group, 2004) and the Directory of Open Access Repositories (OpenDOAR) (Centre for Research Communications, 2013). These registries profile, aggregate and analyse repositories, and study the evolution and adaptation of *open access* from various perspectives. Such efforts require the development of new methodologies and approaches for observation. One of the early efforts was ROAR analytics. Using ROAR analytics, Carr and Brody (2007) proposed an approach for measuring research community engagement within *institutional repositories* that have taken advantage of the large-scale adoption of the interoperability standards (Lagoze and Van de Sompel, 2003), which in turn has enabled the automated machine processing of distributed repositories. This approach is easy to adopt and services can perform the analysis automatically. Consequently, it has been taken by the Registry of Open Access Repository Mandates and Policies (ROARMAP) (ROARMAP, 2010) and by the MELIBEA directory of institutional policies (Vincent-Lamarre et al., 2016). These estimate and measure the strength and success of open access mandate policies, cross-validating them with the analysis of Web of Science datasets and open access repository metadata (Vincent-Lamarre et al., 2016).

These types of practice are distinguishable for the following reasons: I) they follow an open access publishing agenda, and II) they incorporate *data analytics* practices. In this thesis, this form of practice is denoted as ‘open access analytics’. Open access analytics is the process of a systematic examination of data to gather insight about an open access publishing agenda. Indeed, open access analytics is a critical supplement to the advocacy and development of open access publishing. This is due to the debate raised by the research community about its development, coverage (Pinfield et al., 2014) and impact (Harnad and Brody, 2004) on the scholarly publishing system.

However, data analytics is highly dependent on the availability of data about the subject being analysed. Thus, the availability of open access repositories and their infrastructure on the Internet presents opportunities to analyse data to understand open access publishing related phenomena, presenting a valuable source that supports open access analytics practices. Furthermore, data analytics using open access repository data is an established practice in the open access research community. According to Harnad (2008b) and Brody et al. (2007), the availability of open access materials in open access repositories means they also work as scholarly databases, as quantitative data can be generated from open access repository data. In addition, they highlight three requirements to achieve high coverage: functionality, intensives and mandate.

In a way, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) framework is the functionality that supports their vision. The OAI-PMH is a low-barrier interoperability protocol that achieves interoperability by means of metadata harvesting (Lagoze and Van de Sompel, 2003). Currently, the OAI-PMH is the de-facto standard for open access repository interoperability. Therefore, it is harnessed to provide analytical functionalities, utilising the conventional service and data providers approach. For example, OpenAIRE (Rettberg and Schmidt, 2012) and BASE (Lossau et al., 2006) are directed to provide an open access literature search functionality, which also extends to provide analytical services. Yet, with regard to the COnnecting REpositories (CORE) project (Knoth and Zdrahal, 2012), analytics is a core aim.

However, the strategy of conventional service providers is associated with data providers' and service providers concerns, regarding the improper use of the harvested metadata and resources, as well as bandwidth issues that affect their services (Ferros et al., 2008). Therefore, they adopt data policies, and metadata describes the copyright status of their data (Gadd et al., 2004). Given the fact the number of data providers is increasing, and significant data requires harvesting, this may lead to big data being associated with data management challenges and data acquisition costs (Xia et al., 2017a). From a design perspective, service providers' centralised approach is based on a monolith design. This design is associated with a single point of failure, high set-up costs, and data management and access being limited to a single administrative domain. Thus, it may bring some open data barriers, as it is collected by a party specifically for its own purposes, without any coordination with the users of the collected data (Zuiderwijk et al., 2012). Such challenges act as barriers to exploit the analytical value of open access repository data.

In this doctoral thesis, these challenges and requirements are emphasised, although it also brings attention to a new requirement: the development of analytics methodologies and tools that support analytics practices. On one hand, the current approach of developing these methodologies in the open access community is centralised and specific to a particular administration domain or one time analysis carried out for the purpose of particular research. On the other hand, the Web science research community have highlighted the importance of collaboration in the development of analytics methodologies and activities to understand and analyse the Web, proposing a new decentralised platform to facilitate collaboration and engagement among researchers from different disciplines under the term *Web Observatory* (Tiropanis et al., 2014a). Accordingly, this thesis aims to investigate, encourage and incorporate these efforts in open access analytics and open access repository domains to establish a better understanding of how this form of collaboration can facilitate open access analytics in an open access repository ecology.

1.2 Research Questions

This work has one overarching aim, as well as a number of sub-aims that were developed, during the research process. The overarching aim is to establish an understanding of conducting open access analytics using open access repository data. Within this broad aim, three sub-aims have been addressed to guide three qualitative studies, which are reported in three separate chapters of this thesis. Each of these sub-aims is associated with one main research question. Table 1.1 illustrates the research aims and their associated research questions, as well as the chapters reporting the studies conducted in light of each aim.

Study Aims	Research Questions (RQs)	Chapter Title	Ch. No
To establish an understanding of the delivery of open access analytics using open access repository data, by taking a conventional OAI-PMH service provider approach.	To what extent does the OAI-PMH service provider conventional approach provide adequate support to operate open access analytics using open access repository data?	Open Access Analytics Using the OAI-PMH Service Provider: ROAR Analytics (Case Study)	4
To establish an understanding of how the concept of *social machines* can support the process of open access analytics.	Is the concept of social machines useful to reconceptualise the process of open access analytics within an open access repository ecology?	The OAA-OARD-SM Conceptual Framework Development	5
To establish an understanding of open access repository data exploitation for analytical practices at the repository level.	How is open access repository data exploited for analytics practices in the UK based on open access repositories by their management?	Open Access Repository Management Interviews	6

Table 1.1: The research aims and questions and their associated chapters in the body of this thesis.

1.3 Research Contributions

This thesis contributes to open science research literature by providing an understanding of open access analytics practices. It begins by demonstrating open access analytics using the data service provider approach that is facilitated by the de-facto interoperability standard, which is used to interoperate repositories and exchange data between open access repositories. Following that, it introduces a new conceptual understanding, provided as a conceptual framework, of open access analytics using open access repositories with the social machine concept. In addition, it establishes an understanding of analytics practices using open access repositories data within the repository and its management. These three main elements contribute to providing a better understanding of open access analytics at the service provider level (case study), the process level (conceptual framing using conceptual analysis and conceptual synthesis) and the open access repository level (analytical exploitation of repository data based on the perspective of open access repository management as an expert of open access repository domains).

1.4 Thesis Structure

This thesis is divided into seven chapters, including this introductory chapter. **The introductory chapter** provides an overview of the research background and motivations, as well as introduces the reader to the researcher's research contribution and the research question that the research derives from, including its aims and objectives. The remaining chapters are as follows:

Chapter 2: The background research and literature review. This chapter aims to provide a fundamental understanding of the research related concepts, including a synthesis of the open access analytics concept, open access repository and open access repository data relevant concepts and the concept of the social machines, which is associated with the Web Observatory literature. In Chapter 2, the discussion of the thesis' concepts is associated with their role in this thesis as well as the adopted perspectives that are consistent with the uses of the concepts across the thesis chapters.

Chapter 3: The research methodology. This chapter provides a theoretical discussion on the research methodology adopted by the researcher to account for the aforementioned research questions. Chapter 3 acts as a theoretical research methodology framework that underpins the research procedures and justifies the choice of research approach as well as the methods used in the research design.

Chapter 4: The ROAR case study. This chapter outlines the first study conducted to answer the first research question. The chapter outlines the case study design, as well as provides an analysis and discussion of the open access analytics delivery process, taking advantage of the ROAR case study to better understand the delivery of open access analytics. The case study design takes the form of an exploratory research design to provide an explanation of how open access analytics can be delivered using open access repository data, wherein the OAI-PMH service provider approach is the means used to operate the analytics. In addition, the limitations associated with the process are outlined with respect to the agenda and the requirements of ROAR analytics.

Chapter 5: The OAA-OARD-SM conceptual framework development. Chapter 5 outlines the research design and process used to answer research question two, which uses conceptual analysis and conceptual synthesis to bring open access analytics and the concept of social machines into a single conceptual framework. The chapter provides the reader with transparency on the conceptual framework development process as one of the requirements of qualitative research. Thus, it goes through the phases of the development as well as presents the justifications for the procedures adopted and decisions taken during the conceptual framework development process. In addition, the chapter provides a narrative and graphical representation of the OAA-OARD-SM conceptual framework.

Chapter 6: The repository management interviews. The conceptual analysis in Chapter 5 draws attention to understating open access analytics from the repository level. Thus, Chapter 6 reports a qualitative study, utilising the repository management team members as experts to establish a better understanding of the analytical exploitation of open access repository data. Furthermore, an understanding is acquired of the delivery model of analytics applications, the role of repository management teams, the coordination and communication with external analytics providers, and the concerns around the analytics practices from the perspective of repository managers as experts in the open access repository domain.

Chapter 7: Conclusion and future research direction. The thesis is concluded by revisiting both the research questions and aims provided in the introductory chapter, as well as identifying the implication of the findings and the understanding acquired as a result of conducting this thesis. In addition, it highlights the future direction of such study and its associated issues that require greater understanding, and calls for more research efforts.

Chapter 2

Literature Review

This chapter provides an understanding of a set of core concepts relevant to the issues investigated in later chapters, including *open access analytics*, *open access repository* and *social machine*. The aim is to achieve a careful review that draws parallels between the concepts of data analytics with open access publishing practices and open access repository as a data source for analytics activities. Similarly, the social machine concept is reviewed, paying attention to its influence on data analytics practices, as it plays a substantial role in the conceptual analysis highlighted in Chapter 5.

2.1 Open Access Analytics

This research limits its scope to investigating the analytics utilized in the open access movement's agenda. Therefore, it is strongly linked with open access publishing, which provides context of the process of data analytics. Accordingly, a basic overview clarifies the concepts related to open access publishing, and the open access repository, as infrastructure supports open access scholarly publishing and the dissemination of content. This overview is significant enough to establish an understanding for later chapters.

2.1.1 Open Access Scholarly Publishing

Scholarly publishing is a circumscribed formal part of scholarly communication, whereas scholarly documents are formally disseminated to be consumed by the scholarly community. Indeed, Lynch (2003) highlights scholarly publishing as a *'very specific'* example of scholarly communication not limited to the dissemination of scholarly output, even though it is specific to the scholarly system that carries out a dissemination role. Kling and McKim (1999) assert that three criteria need to be satisfied by a scholarly document for it to be effectively published within the scholarly community and considered a scholarly publication. These three criteria include the publicity, where the document is registered to its author and made visible to other scholars, the trustworthiness of the document by passing it through a quality control process and the accessibility of the document by its audience in a constant manner over time. These criteria constitute a formal system within the scholarly community, which enables quality control of scholarly knowledge, as well as provides an indicator to evaluate the scholarly community and their output. Indeed, Kling (2004), in his review 'The internet and unrefereed scholarly publishing', stated that:

> Scholarly publishing is one formal part of scholarly communication, and serves as a basis for scholarly evaluation. Scholars and academic programs are often reviewed, in part, based on the quality and quantity of their research published in journals; the quality of journals is often assessed by the 'impact factors,' measured by citation analyses. (Kling, 2004, p 593)

2.1.1.1 Scholarly Journal Publishing

Although, scholarly publishing encompasses varying forms of publishing , including scholarly books, scientific monographs and scholarly journal publishing (Abel et al., 2002), Kling (2004) uses the example of scholarly journal publishing to represent scholarly publishing, as it is the predominant scholarly publishing system. Indeed, it has a legacy of propagating scholarly knowledge spanning four centuries (Swan, 2006) and involves a set of mechanisms which evolved to evaluate and control the formal scholarly knowledge published in scholarly journals (Rush, 1996). In their 50th Anniversary report, the International Association of Technical and Medical Publishers (STM) described scholarly journal publishing as "... periodicals carrying accounts of research written by the investigator themselves and published after due peer review". They add that: "This type of journal publishing [scholarly journal publishing], involving the peer reviewed first reports of phenomena or ideas'...".

These periodicals embody four main functions: the registration, the dissemination, the certification and the archival record of knowledge. Most of the functions embodied are also core to scholarly publishing, as it is, as stated earlier, distinctive due to its quality control system, namely the peer-review system. Peer review is a process of utilising an anonymous expert in a field to ensure the quality of a scholarly paper (Björk et al., 2009). The scholarly paper undergoes a peer-review process to become identified as a scholarly journal paper, which then forms part of a scholarly journal (ibid.).

Scholarly journal publishing is also distinctive due to its economic model. In the post-World War II period, scholarly journal publishing was dominantly founded and controlled by societies and associations until a commercial scholarly publisher managed to establish a scholarly journal on a solidly profitable basis in 1950 (Guédon et al., 2019).

What followed was the creation of the Journal Impact Factor (JIF) index, which created a competitive market that generated 9.4 billion in revenue in 2011 (ibid.). The market was dominantly a subscription-based model that enabled individual scholars to subscribe to a particular journal, before a distinct shift to library-provided materials enabled academic institutions (such as libraries) access to the publisher's collections by acquiring a subscription. Later, a pay-by-the-drink model was introduced to allow scholars to pay for a particular journal article (Cox, 1998).

Another distinctive part of journal publishing is its value chain. The value chain of scholarly journal publishing involves a number of actors, including (but not limited to) the research funder, the institution, the authors, the peer-reviewers, the publisher and libraries as customers. Thus, a substantial part of its value chain takes place in the academic community, whereas other parts are processed by intermediate commercial publishers, including dissemination and indexing services (Björk, 2005). While the publishers are driven by profit, the author, as core actor in this value chain, is driven by a need to be read by their peers and cited (Eve, 2014b).

According to Harnad (1994b) in his 'Subversive Proposal', academics are 'esoteric', motivated by the need to be cited and read by their peers rather than sell their work. What emphasises this situation is the link made between their academic position and the status of their published work. Academics are evaluated and promoted based on the quality of their publishing activities (Eve, 2014a). Hence, on the one hand, authors share their work without royalty, and the academic community carries out the peer-reviewed process for free. On the other hand, the publishers place a paywall over their collection, enabling them to receive a profit margin of 20–30% according to analyst estimation provided in Noorden (2013).

With these changes in acquiring access to commercial-based scholarly journal publishing and the situation the publisher is placed in, two main economic problems arise. First, exclusive control over scholarly publishing is held by the commercial publishers (De Silva and Vance, 2017) due to their dominance. By 2003, three main publishers (Elsevier, Springer, and Taylor and Francis) controlled 60% of ISI Web Science leading citation index journals (Willinsky, 2003). Second, in addition to the dominance of commercial publishers, the price enforced by publishers to access a collection is increased, creating a barrier for academic institutions. This access barrier limits the ability of academic institutions to provide access of the publisher's collections to their audience. According to data from the Association of Research Libraries on the increase of book and journal prices from 1986 to 2006, prices increased by 180% (Linda, 2018). As a result of this, research libraries have been forced to increase their budget or cut off their subscription. Even academic institutions with high incomes and reputations face challenges in funding their access to journals they used to subscribe to in what has become known as *'the serials crisis'* (Swan, 2006) or *"journal affordability problem"* (Harnad et al., 2004).

In addition to the challenges associated with the subscription-based business model, the shift towards digitisation and the evolution of the World Wide Web (WWW) have significantly contributed to the emergence of a new form of scholarly publishing. Traditionally, scholarly journals were disseminated through the mass printing of 'paper journals'. However, as the digital era emerged, scholarly journal publishing began to publish material in a digital format, which has led to the development of 'electronic journals' in addition to traditional paper journals (Kling, 2004). What has followed has been the rise of electronic publishing as the only format in which journals reach their audience (ibid.). Yet, the emergence of the World Wide Web has enabled worldwide access to electronic journals (Swan, 2006).

At the scholarly paper level, new concepts have been introduced to highlight its electronic form, as well as its stage in the peer-review process. The electronic form of a scholarly article is identified as an eprint. Eprints are the digital form of peer-reviewed articles, which are typically classified into two categories: preprint and postprint (Harnad, 2004). The two classes represent two main phases of the peer-reviewed journal article life cycle. According to the NISO/ALPSP JAV technical group, preprints take the form of the author's original copy or submitted manuscript under review, and postprints can be delivered as accepted manuscripts, proof copies or a version of record (Morgan, 2008).

The shift from paper-based publishing to e-journal raises the question of how much it costs to publish a journal paper. Noorden (2013), in his paper 'the true cost of science publishing', highlights the debate in the scholarly community on the actual cost of journal publishing. In a broad sense, the debate is articulated into two main perspectives: the open access supporters, who argue that the current cost is exaggerated, and open access journals provide evidence that the publishing cost can be reduced to a minimum, and the publisher's perspective that publishing is not only limited to distribution and printing costs, but it includes the values of editing provided by a professional editor, indexing and the cost of the paywall system. In addition to the cost of rejection, scholarly journals with high impact factor are highly competitive and involve a large collection of rejections that cost more than journals with low impact factor, which in its turn provides filtering and a high-quality collection to authors. In addition to the actual cost, the scholarly community and the digital economy is affected by the exchange of symbolic capital that is represented in the prestige of a particular journal article. According to Eve (2014a, p 44), prestige is 'a proxy measure for quality that is gained through an economic rationing of material'. Eve (2014a) in his book, Open Access and the Humanities: Contexts, Controversies and the Future , highlights the cycle of prestige in scholarly publishing, stating that:

> The acceptance of such research by publishers who possess both material capital (needed to undertake the labour and effectively disseminate the work) and cultural capital (knowledge of publishing and academic systems) constitutes a payoff in the form of social capital (endorsement and support) for the

> author that can be re-converted back into the symbolic capital (prestige/reputation) that is needed for peer respect and a job/ promotion (material capital). Acquiring authors with high levels of cultural, social or symbolic capital for their list improves a press's own social, symbolic and material capital (in the ability to sell research). (Eve, 2014a, p 45)

This flow highlights how the incentives to pursue prestige are created and, consequently, how the value of prestige is constituted in the scholarly community, even though this value can be limited and constrained by the paywall instead of being an excuse to its existence. Eve (2014a) argues that prestige can support the dissemination of scholarly material, even though it can be limited to the discoverability of the scholarly journal article due to the existence of the paywall.

2.1.1.2 The Rise of Open Access Scholarly Publishing

This mix of economical, technical and social factors contributed to a substantial shift in the scholarly journal publishing paradigm. It led to the rise of a new social movement within the scholarly community that called for an immediate removal of any barriers preventing access to scholarly journal publishing known as the 'open access movement', which has recently emerged under the broader umbrella of what is referred to as 'open science'.

Fecher and Friesike (2014) examined open science as a concept within scholarly communication literature. They identified five main open science schools of thought based on the assumptions made by their research community. These five schools of thought include the infrastructure school (which calls for innovation in open technologies and infrastructure to support the research process), the public school (which calls for the public to have access to knowledge creation), the measurement school (which calls for alternative impact measurement from peer-review), the democratic school (which calls for free access to knowledge) and the pragmatic school (which calls for collaboration between researchers and the opening up of the process of knowledge creation). In this section, the primary focus is the Open Access Movement (which can be identified as part of the democratic school of thought) in the context of analytics (which can be linked to the infrastructure school).

2.1.1.3 The Concept of Open Access

Open access is an open science movement (Pontika et al., 2015)) concerned with making peer-reviewed (postprint) and unreviewed (preprint) articles freely accessible on the public Internet (Suber, 2007). As highlighted above, its emergence strongly correlates with the evolution of the World Wide Web. However, in practice, it is rooted in the establishment of computer networks , since it began with computer scientists exchanging electronic forms of their publications via email, thus making it free and easy to access scholarly papers (Swan, 2006). However, its initial milestone as a movement was the establishment of the first scientific archive in 1991, followed by Harnad's 'Subversive Proposal' to adopt online open access publishing (Harnad, 1995b).

Yet, the official announcement of open access as an initiative was declared by the Open Society Institute in Budapest in 2001 with the Budapest Open Access Initiative declaration defining open access as

> By "open access" to the literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. (BOAI, 2002)

BOAI (2002)'s definition asserts that the public Internet is a medium that can give open access to research, resulting in it not being specific to a particular community or geographical area, but providing free global access to scientific research using Internet access as a minimum requirement. Based on this definition, open access is also not regarded as being specific in its purposes. The definition goes beyond granting free access rights. Instead, it grants the right to distribution and reuse, as well as the right to access the research. Furthermore, although the barriers of price and copyright are removed, authors still retain the right to be acknowledged and cited.

2.1.1.4 The Scope of Open Access Movement

Although the ultimate aim of open access is to make research more openly accessible, the BOAI (2002)'s declaration defines the scope of this form of scholarly publishing as having to ensure it is openly accessible:

> The literature that should be freely accessible online is that which scholars give to the world without expectation of payment. Primarily, this category encompasses their peer-reviewed journal articles, but it also includes any unreviewed preprints that they might wish to put online for comment or to alert colleagues to important research findings. There are many degrees and kinds of wider and easier access to this literature.(BOAI, 2002)

As highlighted in the BOAI declaration, the open access movement limits its target to scholarly journal articles only. In addition, the statement emphasises one of the distinctive natures of scholarly journal publishing highlighted in Section 2.1: the fact that scholarly papers are provided by the authors without receiving any revenue, with the primary target of publishing focusing on the impact of their research instead of payment to the research publication itself. Since authors are not paid for most scholarly journal articles, the OA movement considers them to be a prime target of those that should be made freely accessible (Bailey, 2006).

Harnad (2004, 1995a) makes a distinction between 'give-away' literature and 'non-giveaway' literature, and the open access movement aims to make give-away literature openly accessible. He also determines solely one criterion to distinguish between the two types of literature, specifically, whether the authors seek any royalty or fee in exchange for their writing. In the case of give-away literature, authors are focused on the impact of the research and give away their text without asking for any fees or royalties in return. However, non-give-away literature is written by authors seeking revenue in return for publishing their text.

Suber (2012) states that the open access movement was created by researchers aiming to overwrite the restrictions made by publishing institutions in order to increase the impact of their research. Yet, the declaration by BOAI includes a call for both preprints as well as postprints to be made openly accessible to all. Thus, the target is for open access to be extended to unreviewed and peer-reviewed journal articles, provided in the forms of both preprint and postprint.

2.1.1.5 The Green Open Access vs the Gold Open Access

The BOAI (2002) declaration describes two main strategic routes to make research open access. The first route is a journal-level strategy, known as 'gold open access' (Suber, 2007). It is mainly based on creating new peer-reviewed journals with no access barriers or converting a subscription-based journal into an open access journal. The second strategic route is referred to as the 'green open access'. This is an author-level strategy, which involves authors contributing their work by depositing it in open access repositories established by academic institutions or disciplinary bases, known as an open access repository (Harnad, 2006).

The types of journals that use the 'gold open access' strategy include pure gold open-access journals, hybrid open-access journals and delayed open-access journals. The pure gold open-access journals provide all their scholarly articles openly accessible, as their income is based on alternative funding options rather than subscription fees, including (but not limited to) article processing charges (APC), which charge authors' funders or apply one-time fees per author. In the case of gold open-access journals, making the scholarly journal article open access is not an option, instead it is compulsory, as the journal is fully open access and does not apply any price barrier on top of their collections. However, hybrid open-access journals adopt both the subscription-model and open access publishing model, as they apply a paywall on their articles, yet the authors pay APC charges to make their research open access. While both pure gold open-access journals and hybrid open-access journals enable immediate open access to research output, the delayed open-access journals transform their collection into open access a period after the publishing date (Martín-Martín et al., 2018).

Lastly, green open access adopts open access repositories as platform pools where an author can deposit his/her paper and offer it to the research community as in an open access format. The open access repository is a core concept in this research, therefore in-depth discussion is provided in Section 2.2.

2.1.1.6 Summary

This section introduces the reader to basic concepts correlated to open access publishing, including scholarly publishing, scholarly journal publishing, scholarly journal articles, open science and open access publishing. It also distinguishes between the various assumptions about open science and highlights routes to open access. More importantly, it directs the reader from a broad concept to very specific concept that determines the scope of this research. This research is concerned with open access publishing as the context of targeting the green open access route to harness its infrastructure and examine the opportunities for analytical application of the infrastructure. The following section provides a theoretical umbrella, concerning data analytics, in order to integrate it with this section to define the concept of open access analytics.

2.1.2 Data Analytics

Analytics is a relatively new term and an overarching concept that intersects with various research areas, such as operational research, statistical analysis, data-driven decision-making and knowledge discovery (Cooper et al., 2012). According to Rose (2016), the term 'analytics' was first introduced by Davenport et al. (2006) in their research report "Competing on Analytics" in May 2006. Since then, the term has become widely used, although it has created some confusion regarding its meaning, as well as its connection with analogue terminology. Rose (2016) added that the term 'analytics' is used in three different ways, specifically, as a synonym for statistics and matrices, a synonym for data science and a general term, representing the quantitative approach to organisational decision-making. Yet the use of the term, when referring to quantitative analysis and decision science, can be considered its broader use. Nevertheless, Hawkins (2008) argues that analytics is a goal-oriented practice, and therefore several terms are associated with the term 'Analytics', including 'Academic Analytics', 'Learning Analytics' and 'Business Analytics'.

Both the rise of sub-concepts and the linkage of data analytics with other broader concepts has enabled several research communities to contribute to analytics research. For example, Cooper (2012a) identified seven research communities that contribute to analytics research, including statistics, business intelligence, web analytics, operational research, artificial intelligence and data mining, social network analysis and information visualisation. Moreover, a visual analytics research community has emerged from information visualisation and data mining research (Keim et al., 2008).

2.1.2.1 The Concept of Data Analytics

As each of these research communities creates their own concept and definition of this term, there is an increasing absence of agreement on a definition for data analytics. In addition, the analytics research community is more concerned about the analytical infrastructure and their systems, instead of addressing analytics as a process beyond its technological complexity (Cooper, 2012a). Despite the fact that several definitions are provided (Table 2.1), a consensus on the essence of the concept of analytics exists, as the majority of the definitions clearly state that analytics is a process carried out for the purpose of supporting problem-solving and decision-making.

One of the sub-aims of this research is the establishment of an understanding on how the concept of social machines can support the process of open access analytics, whereas the analytics process is central to understanding social machines (see Section 2.3.1). Therefore, it is important to conceptualise analytics as a process that adopts a definition that aligns with this perspective. Three out of five definitions highlighted in Table 2.3 clearly define analytics as a process.

Keenan et al. (2018) and INFORMS (n.d.) definitions characterise analytics as a process that involves a set of data analysis activities to achieve better decision-making. On the other hand, Cooper et al. (2012)'s definition characterises the output of analytics as *actionable insight*, instead of better decision-making. Also, it elaborates on process activities, generalising it into two main phases, namely problem definition and the application of analysis methods. This can ease the complexity of conceptualising the data analytics process, excluding decision-making complexity from the data analytics process, and determine a concrete output of the data analytics process. Nevertheless, there is considerable debate on what constitutes *'insight'* in data analytics. Therefore, a further discussion on the definition of insight is provided in following section.

In addition to conceptualising analytics as a process, Cooper (2012b)'s and Keenan et al. (2018)'s definitions are accompanied with conceptual frameworks that characterise the concept of analytics. Keenan et al. (2018) provide a conceptual framework which classifies analytics into two main approaches; *data-centric* approaches and *decision-centric* approaches (see Section 2.1.2.3). However, in his article, "A Framework of Characteristics for Analytics", Cooper (2012b) provides an insightful conceptual framework that provides an anatomy of analytics, which has been partially utilised in this thesis to define ‘open access analytics’ in Section 2.1.3.

To sum up, this thesis adopts Cooper et al. (2012)’s definition that defines analytics as

> the process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/ or simulated future data. (Cooper et al., 2012, p 3)

The rationale for adopting this definition is to enable conceptual synthesis with the social machine concept to place special attention on the process within a particular social machine, to minimise the complexity of conceptual analysis of the analytics process carried out in Chapter 5 by adopting insight as the output of the analytics process and to take advantage of the accompanying conceptual framework, which provides an anatomy of analytics that facilitates the conceptual analysis of the open access analytics concept in Section 2.1.3.

Definition	Reference	Remark
"a process by which a team of people helps an organisation make better decisions (the objective) through the analysis of data (the activity) "	Keenan et al. (2018, p 2)	It includes any form of data analysis and highlights the analytics as team work activity driven by decision support requirement within a organisation.
"extensive use of data, statistical and quantitative analysis, exploratory and predictive models, and fact-based management to drive decisions and actions. The analytics may be input for human decisions or drive fully automated decisions"	Davenport Thomas and Harris (2007, p 7)	It highlight analytics as a management paradigm associated with extensive use of data and analysis methods and outline both analytics as support to human cognition and the prescription analytics, whereas the decision process take place in automated form.
"analytics is the scientific process of transforming data into insight for making better decisions"	INFORMS (n.d.)	It highlights the output of analytics that can support the decision-making process. In addition, it denotes analytics as a *"scientific process"* which implies the systematic nature of the analytic process.
"the process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/ or simulated future data."	Cooper et al. (2012, p 3)	In addition to the emphasis on the analytics as process, Cooper et al. (2012) draw attention to the primitives of analytics; the problem definition and the usage of analytics methods as well as constrain the insight to the *"actionable insight"*.
"the application of processes and techniques that transform raw data into meaningful information to improve decision making"	Wilder and Ozgur (2015, p 180)	They define the analytics as transformation process transform the data into meaningful information (knowledge).

Table 2.1: A summary of definitions of analytics

2.1.2.2 Insight as Output of Analytics

The actionable insight in the context of Cooper et al. (2012) 's definition is the result of the analytic process that is beyond theoretical description or mere reporting, rather it is associated with a potential and practical action. However, insight is not always associated with action, but instead is considered a unit of discovery. Saraiya et al. (2005, p 444) define insight as "an individual observation about the data by the participant, a unit of discovery".

Based on Yi et al. (2008)'s study, which surveyed the approaches used in Information Visualisation to gain insight, Chang et al. (2009) concludes that the uses of insight are more or less a unit of knowledge. However, they argue that considering insight as knowledge or information limits the scope of analytics to become a knowledge representation process, which is not the case in insight gathering. To avoid the insight definition dilemma and highlight insight complexity, North (2006) highlights five main characteristics of insight, as shown in Table 2.2.

Characteristics	Description
Complex	It more complex than a single value and introduced from a large amount of data.
Deep	It is cumulative of insight gathering processes and developed over time. Therefore, according to Sacha et al. (2014) model, the insight may lead to knowledge generation or hypotheses and consequently new insight about the data.
Qualitative	It is not quantitative, as it involves subjectivity, and therefore can be uncertain.
Unexpected	It is often unpredictable.
Relevant	It is deeply linked with the data knowledge domain.

Table 2.2: Characteristics of insight adopted from North (2006)

These characteristics imply how complicated the insight gathering process is, as it often cannot be achieved within a predefined task. Therefore, the human element should form part of the process, instead of reproducing the results using apps with a predefined data analysis methodology. Nevertheless, the use of these apps supports the insight gathering process and reveals hidden patterns in the data.

2.1.2.3 Analytics, Knowledge Discovery, Information Visualisation and Visual Analytics

Based on the tasks analytics is used for, analytics can be classified into three categories: *descriptive, predictive and prescriptive.* While descriptive analytics is based on reporting the data to understand and gather insight using human cognition with the support of the analytics system, predictive analytics utilises data-mining logarithms to produce a statistical model that can be used to deduce knowledge not implicitly provided in the data. However, prescriptive analytics takes it a step further, enabling the machine to offer action choices or acts based on a simulation or decision model (Cooper, 2012a; Delen and Demirkan, 2013). Consequently, the analytic process can take the forms of an automated task using data-mining logarithms or a human cognitive process that is supported by analytic systems and tools (Grolemund and Wickham, 2014)

Another form of classification is determined by the approach used to carry out the analytics process, as provided by Wegryn (2014) and Keenan et al. (2018). According to Wegryn (2014), analytics can be broadly classified into data-centric analytics and decision-centric analytics. Data-centric analytics is based on the exploitation of the value of available data in order to establish any interesting insights that can help to anticipate a particular issue or problem. Thus, the analytics process starts from the data. However, decision-centric analytics is driven by the need to understand a particular problem in order to solve it. Thus, the requirement of the process is determined by what is required to understand the problem at hand. To clarify, the process starts with understanding the problem related to a particular decision, then it defines the objectives of the analytics and ends with the making of a decision. In contrast, data-centric analytics is associated with a generic goal and takes the form of exploratory analysis. Knowledge discovery and data mining applications are well-known examples of data-centric analytics. Knowledge discovery is "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". (Fayyad et al., 1996, p 30).

Thus, the analytics process is based on the availability of the data and the exploration of this data to find a useful pattern. Fayyad et al. (1996) present the data mining process as an integral part of the whole knowledge discovery process, which is utilised to uncover hidden patterns from a large data set. They define data mining as a sub-task of the knowledge discovery process, which consists of applying data analysis and discovery logarithms to data to produce a particular pattern.

Information Visualisation (InfoVis) is a substantial component of the analytics process and supports all forms of analytics, that is, data-centric and decision-centric, as well as descriptive, predictive and prescriptive, as its ultimate aim is to amplify human cognition and support the understanding of the data and the insights collected during the analytics process. Card et al. (1999) argues that InfoVis can amplify human cognition by increasing the processed resources, reducing the search by representing a large amount of data in a small space or by enhancing the pattern recognition process. Similarly, distributed cognition researchers have argued that human cognition is not limited to the human mind only but is distributed among the people and artefacts around them, influenced by their environment. Therefore, the visualisation is presented as an external tool that should be harmonised with human internal cognition (Hutchins, 1995; Clark and Chalmers, 1998; Liu and Stasko, 2010).

Card et al. (1999, p 7) define InfoVis as “the use of computer-support, interactive, visual representation of abstract data in order to amplify cognition”. To clarify, InfoVis is about creating effective mapping between data and visual representations to support the process of making sense of the data. However, it is not limited to visual representation, but the interaction between analytics and visual representation is a core aspect of Information Visualisation.

Another concept related to data analytics, and outgrowth from the field of InfoVis, is visual analytics (Wong and Thomas, 2004). Visual analytics (VA) represents the integration of visual and automated analysis facilitated by data mining, machine-learning methods and human interaction with computerised visualisation. According to Thomas (2005, p 10), VA is “the science of analytical reasoning facilitated by interactive visual interfaces”.

Thus, using the VA concept, the analytics process is investigated from the exploration of data to reasoning and insight gathering. A set of process models are provided by the VA research community, which integrates the knowledge discovery process and the InfoVis process, where the analytics process is illustrated as an iteration process that takes place in the computerised system and human cognition is facilitated by user interaction and system processes to generate models and visualisations (A further discussion is provided in Chapter 5).

2.1.3 Open Access Analytics: A Conceptual Analysis

To define open access analytics, this research makes use of two concepts contributed by Cooper (2012b): *the analytics subject* and *the analytics object*. While the analytics object is the entity affected by the use of analytics and the result of made decisions, the analytics subject is the entity that the data is about. Therefore, the term ‘open access analytics’ refers to a data analytics process characterised by open access publishing, where the analytics object is open access publishing despite the analytics subject being used in the analytics process.

In the open access research community, analytics is a substantial component that supports several agendas, including, but not limited to, open access adoption analysis (the analysis of the quantification of open access growth, repository adoption and open access journal adoption), open access policy monitoring and evaluation (the monitoring of adopted policies that support open access publishing as well as the evaluation of their effectiveness) and open access advantage analysis.

In addition, open access resources have been used to carry out the analytics process for various agendas apart from the open access publishing-related agenda. For example, Harnad (2008a) draws attention to the fact that open access literature provides opportunities for scientometric practices and presents it as an alternative source of data to monopolised commercial scholarly databases, such as Web of Science (WoS) and Scopus databases, as well as limited access databases, such as Google Scholar, provided by Publish or Perish software tools. While this form of opportunities motivates the research community to adopt open access publishing and the data used in the analytics process forms part of the open access literature, the analytics object is not open access publishing itself. Thus, it is out of the scope of open access analytics.

2.2 Open Access Analytics (OAA) using Open Access Repository Data

Understanding open access analytics as a data analytics process characterised by open access publishing as the analytics object enables various possible data sources to carry out open access analytics, including (but not limited to) open access journal webpages, subject-based repositories, institutional repositories, Google Scholar's data, Microsoft Academic data and commercial scholarly databases. Open access repositories are one of the data sources that can provide rich data on open access publishing and hold representative data according to green open access publishing. Therefore, this section provides an overview on open access repositories and open access repository data before it is integrated with open access analytics in Section 2.2.3.

2.2.1 Open Access Repository

The term open access repository (OAR) can be regarded from two main perspectives: the digital preservation research-community perspective or the scholarly communication research-community perspective. Neither of these indicate the openness of the repositories. However, they define the repository based on its structure and the role it is committed to fulfil.

2.2.1.1 Open Archive, Trusted Repository and Digital Repository

From the digital preservation research-community perspective, two terms are used: the open archive and the trusted repository. According to the CCSDS (2012) guide on the Open Archiving Information System (OAIS) reference model, the open archive is an archive that is responsible for the preservation of digital resources used by a designated community. Furthermore, the designated community is the producers and consumers of the digital resources, and the term 'open archive' refers to an intermediary entity that preserves the digital resources. However, the trusted repository is an organisational entity established for the role of providing its community with 'reliable' and 'long-term access' to managed digital resources on behalf of their depositor (RLG-OCLC, 2002).

While the two definitions are consistent to a wide extent, a trusted repository is introduced to encompass the repositories founded to preserve scholarly materials. Thus, the concept should reflect the complexity associated with the preservation of information by the scholarly community. This presents the open archive as a higher level of abstraction of repository implementations, and the trusted repository concept is compatible to the open archive concept associated with the adoption of an open archive in the scholarly community (RLG-OCLC, 2002).

Koutsomitropoulos et al. (2004) use the concept of digital repository, where the concept is meant to be inclusive of both the OAIS concept (CCSDS, 2012) and the trusted repository concept (RLG-OCLC, 2002), as both can be described as a digital repository. Koutsomitropoulos et al. (2004, p 272) provides a broad new definition to represent the fact that it has a wide variety of purposes and uses, defining a digital repository as "a collection of digital entities that are subject to the following three operations: insertion, deletion and retrieval".

Across these three views, the ontology of a repository varies between a collection of digital entities, an organisational entity and an archive. Based on the perspective of each community about the role of the repository, Brody (2006) points out that both an archive and a repository are used in the literature of digital libraries and the open access research community. While the literature found in a digital library highlights that the repository is an entity with a preservation role, the open access community focuses mainly on the dissemination and the access to the literature. Alternatively, Koutsomitropoulos et al. (2004) presents another definition to overcome the lack of agreement on a common definition, utilising the structure and basic function of repository.

On the other hand, the perspectives of those in both the scholarly community and the open access community agree with the view that the repository is an organisational entity. Lynch (2003) defines the repository in the context of an institutional repository, stating it to be:

> a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. (Lynch, 2003, p 328)

He adds that

> It is most essentially an organisational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organisation and access or distribution. (ibid.)

Thus, the repository is a service that represents an organisational commitment of the institution to provide for its community. He emphasises the paradigmatic shift made in the scholarly communication system of using repositories to act as stewards of the digital material, as well as taking a preservation, as well as a dissemination role.

2.2.1.2 Open Access Repository: Institution and Subject-Based

The open access repository concept is not limited to a specific theme or purpose of repository. Instead it is related to the compliance of the repository to provide open access, as clarified earlier. In a broad sense, an open access repository is a set of systems and services that provide the open access community with operational services to manage, retrieve, display and reuse open access resources (Pinfield, 2009). Repositories are developed for various purposes, stewarding different types of electronic materials, including, but not limited to, e-learning repositories, data repositories, thesis repositories and eprint repositories. Two main types are distinguished in scholarly communication literature: an Institutional Repository (IR) and a Disciplinary or Subject-Based Repository (SBR) (Armbruster and Romary, 2010).

While IRs are established by academic institutions or research funders, SBRs are founded by community members. In addition, SBRs target broader community output, and their value is more focused on communication and the advantages of dissemination. They are theoretically more well-defined, as they focus on a particular subject or research area (Armbruster and Romary, 2010). However, IRs can capture the intellectual capital of their institution by organising open access literature into disciplinary themes (Kim, 2007). Therefore, their content is not limited to peer-review or unreviewed materials. Indeed, they host original institutionally produced digital material, including, but not limited to, educational materials, research data and eprints (Crow, 2002).

Associating IRs with the role of capturing institutional intellectual capital enables parallels to be drawn between them and Current Research Information Systems (CRIS), which are databases established to manage institutional research, including funds, human resources and research activities (De Castro, 2019). Although IRs are distinctive due to their set of features, De Castro (2014) highlights seven characteristics that distinguish the institutional repositories from CRIS. According to De Castro (2014), an IR is an open, full-text and externally-oriented repository, dedicated to disseminating research output. Therefore, IRs are dissimilar to scholarly open access publishing in terms of their public availability. However, CRIS are founded for institutional research management.

To sum up, the open access repository is a concept that incorporates repositories that comply with the concept of open access and provide open access materials. Therefore, both IRs and SBRs can be considered open access repositories, as they host eprints and provide permanent open access to their resources. However, this thesis focuses on

IRs, as they are open access repositories associated with a well-formed community and operated by an academic institution.

2.2.2 Open Access Repository Data

To understand what open access repository data is, it is important to review the concept of data. The term *'data'* is used to describe a set of values that are raw and exist in any form. In the context of knowledge discovery, Fayyad et al. (1996, p 41) define data as "a set of facts", such as those stored in databases that can be without any meaningful relation or pattern. Similarly, Ackoff (1989) describes data as symbols that represent properties of an object, event or their environment, which are the products of observation. Thus, in general, open access repository data is a set of values stored and managed in the repository system.

To review the concept of *open access repository data* more closely, three concepts are discussed, including scholarly data, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) data and institutional digital content. These concepts were determined based on three common properties associated with open access repository: I) an open access repository that operates in the scholarly community and hosts scholarly data , II) interoperability is one of the key aspects of the open access repository and the de facto standard for interoperability in an open access repository is the OAI-PMH protocol and III) the domain of the open access repository content expands to include the varied forms of digital content produced within an academic institution.

2.2.2.1 Digital Content

Open access repository content varies according to the scope and the type of repository. For example, the content domain of an IR expands to encompass the 'intellectual product' of an institution, instead of merely being dedicated to scholarly publishing material as suggested by Harnad (2001). Crow (2002) describes an IR as "the digital collections capturing and preserving the intellectual output of a single or multi-university community". Similarly, Lynch (2003) argues that the content domain of an IR can be any digital materials created by the institution and its members. Therefore, IR content may include research articles, research data, educational resources and digital material, such as audio or video files (Genoni, 2004).

Markey et al. (2007) examined the different aspects of an IR in a census covering 446 academic library directors and senior library administrators in the USA, including the document types hosted and managed in the repository. They concluded that a wide variety of documents are hosted in the repositories: theses and dissertations, preprints and learning objects prepared by the institution staff accounting for the majority of repository contents. In addition, the content is not limited to full texts but includes software and any other digital objects produced by the institution and requiring preservation.

This indicates that any digital content produced and used within an academic institution is a candidate for repository content. Conway (2008) modelled the digital-content landscape of universities by reviewing IRs as a digital-asset management service, along with digital collection models. The digital-asset management perspective argues that a repository's content domain encompasses the management of any form of digital content. However, the digital collection model attempts to prioritise the digital content produced within the institution. For example, the collection grid model, presented by the Online Computer Learning Centre (Lavoie, 2003), prioritises digital content based on the level of stewardship required to preserve it and the uniqueness of the content, as illustrated in Figure 2.1.

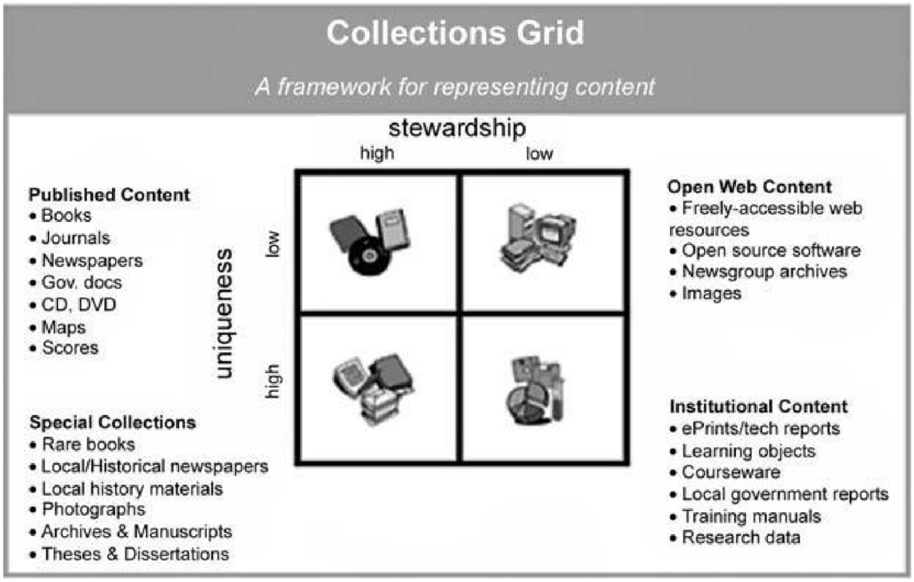


FIGURE 2.1: Collection Grid Model (Lavoie, 2003)

Based on the aforementioned review that harmonises the two perspectives, Conway (2008) proposes that the content landscape model in universities encompasses four types of content. First, e-research includes any data or assets produced during research, such as research data or software. Second, e-teaching encompasses learning objects created for teaching purposes. Third, e-records cover any managed content within the institution, including the institutional repository and the current research information system database. Last, e-publishing refers to content related to scholarly publishing, such as e-journals and books. The Conway (2008) content landscape model provides a broad view of the heterogeneity and complexity of open access repository data.

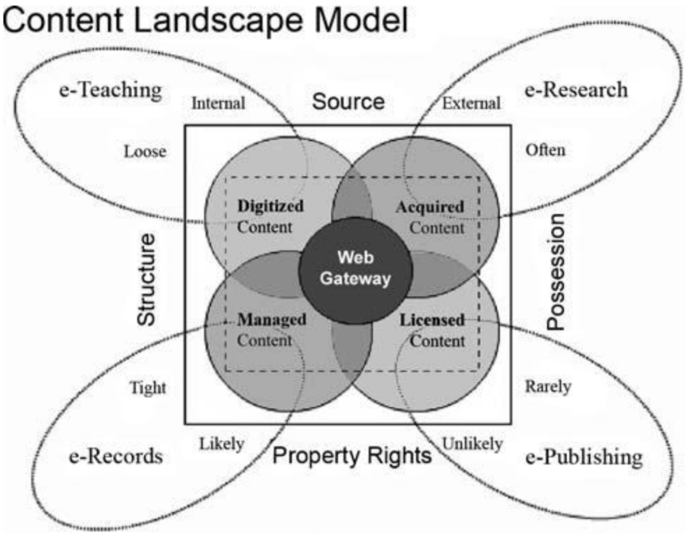


FIGURE 2.2: Content landscape model adopted from Conway (2008)

2.2.2.2 Metadata and Resources

As the vast majority of open access repositories adopt the OAI-PMH protocol as a de facto protocol for interoperability, another useful conceptualisation for open access repository data presents itself. The OAI-PMH interoperability protocol distinguishes between two concepts: metadata and resources. Resources refer to the actual objects, whether physical or digital objects, which are identified by the metadata. Thus, they are not specific to a technological format and not limited to digital objects stored in the repository; instead, it is a concept that accommodates both physical and digital objects, including the type of digital content discussed earlier (Lagoze and Van de Sompel, 2003).

However, metadata is data hosted in the repository. As part of the OAI-PMH protocol, metadata is divided into items and records. An item is an integral component of the repository, which identifies metadata about a particular resource. Thus, it is an abstract element that links the metadata to its resources, and resource metadata is provided as a record associated with its data schema (ibid.). To sum up, OAI-PMH consolidation simplifies the presentation of open access repository data into resources, and metadata is associated with various metadata schemes composed of a set of items and records.

2.2.2.3 Scholarly Data

Despite considerable use of the term *'scholarly data'*, this concept lacks a robust definition. However, a general description is provided by Williams et al. (2014). Williams et al. (2014) describe the scholarly data as

> the vast quantity of data that is related to the scholarly undertaking, such as journal articles, conference proceedings, theses, books, patents, presentation slides and experimental data. (Williams et al., 2014, p 68)

However, Williams et al. (2014)'s definition is not consistent with the concept used in the literature, as often journal articles, conference proceedings, theses and books are categorised as scholarly published materials, rather than being considered scholarly undertakings. A scholarly undertaking can include research and experimental data and is not limited to data extraction or related to the scholarly publishing of materials. Nevertheless, Xia et al. (2017a), Williams et al. (2014) and Xia et al. (2017b) use the term 'scholarly data' to refer to big data in scholarly communication. Their studies are associated with the use of scholarly databases, such as Google Scholar and Microsoft Academic Search, which are examples of databases that contain scholarly data. This implies that the experimental data and research data produced in the research process are not considered scholarly data. Thus, for this research, scholarly data is defined as data that is related to scholarly publishing, rather than referring to scholarly communication in general or scholarly undertaking.

The concept of scholarly data blends those of open access repository data with the vast amount of data produced by the scholarly publishing system. Thus, the efforts made, and the attempt to harness scholarly data for the purposes of analytics, contribute to our understanding of the opportunities offered by data from open access repositories and the challenges that emerge when using such applications.

2.2.2.4 Usage Data

In addition to the concepts relevant to open access repository content, the concept of usage data refers to data that captures the open access repository users' interactions, including downloads, views and system process interactions. In addition, repository platforms record the user's download activities and track the number of times resources are viewed to enable the statistical analysis of collections. Hence, raw data is used to produce useful analytics regarding users' activities and the visibility of repository content (OBrien et al., 2017). In addition, repository platforms also track the system process. For example, a repository platform logs the deposit activity workflow, which is then used to analyse the time spent and effort involved in self-archiving (Carr et al., 2007).

The concept of usage data is broad, and therefore, not specific to open access repository data. Furthermore, it is brought to an open access repository as a result of its role in disseminating open access resources. Hence, there is a need to evaluate its performance based on an in-depth understanding of users' interactions with the usage data as logged by the repository tools.

2.2.3 OAA using Open Access Repository Data: Synthesis

In terms of open access publishing, an open access repository is a key component to achieving the vision of green open access, as stated in the previous section (see Section 2.1.1.5). Therefore, multiple open access repositories must be established to form an open, free, permanent, interoperable and distributed infrastructure that accommodates open access research contributed to by the scholarly community (Lynch, 2003; Awre, 2006). The BOAI (2002) defines an author's deposit of their scholarly paper in an open access repository as the second strategic route to making research open access. Nevertheless, Harnad (2006) argues that the green route should be the main strategy in achieving the goals of the open access movement. With a conservative view of the green route, Guédon (2004) argues that the two routes are complementary strategies, and the community should make use of both simultaneously.

Open access repositories position themselves as entities with the best opportunity of absorbing various forms of scholarly journals after the WWW. Martín-Martín et al. (2018) introduced a model showing the toll-access of journals and open access journals and their position with regard to the paywall in order to illustrate how the Web has the greatest chance of providing free access to academic journal articles. While the Web as a whole is not necessarily compatible with the concept of open access, an open access repository is an open access strategy driven by the open access agenda. Figure 2.3 demonstrates how an open access repository can absorb both toll-access journals and gold open access journal articles.

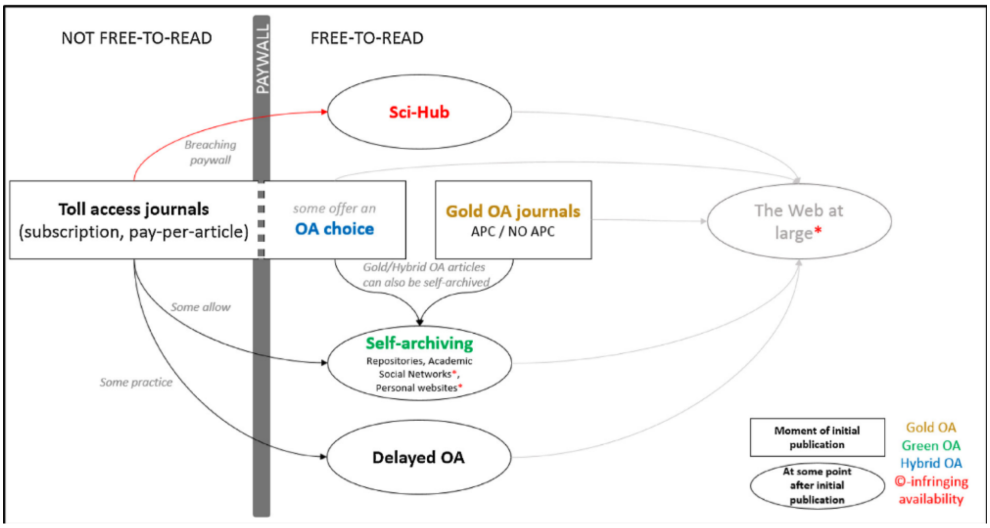

NOT FREE-TO-READ
PAYWALL
FREE-TO-READ
Sci-Hub
Breaching paywall
Toll access journals
(subscription, pay-per-article)
some offer an
OA choice
Gold OA journals
APC / NO APC
The Web at large*
Gold/Hybrid OA articles can also be self-archived
Some allow
Some practice
Self-archiving
Repositories, Academic Social Networks*, Personal websites*
Delayed OA
Moment of initial publication
At some point after initial publication
Gold OA
Green OA
Hybrid OA
©-infringing availability

Figure 2.3: Model of free availability of academic journal articles (Martín-Martín et al., 2018)

In addition to hybrid and pure gold open access, which enables self-archived practices, there are many toll-access journals that allow the self-archiving of their articles in an open access repository. Laakso (2014) examined the publishing policies of the 100 largest journal publishers, whose articles totalled 1.1 million in 2010, to analyse their restrictions against self-archiving practices. He found that 80.4% of these articles can be self-archived as an accepted manuscript or publisher version in open access repositories. A year later, in 2011, Poynder (2011) estimated that the ratio of green open access to gold open access was two-thirds of all the peer-reviewed articles published as open access. This positions open access repositories as significant data sources to investigate in order to gain a better understanding of open access publishing. The following section highlights the nature of open access repository data as a data source to operate data analytics services.

Also, it is essential to evoke the concept of open access in this context. The aforementioned discussion on open access repository data considered comprehensive forms of resources not limited to open access compliant resources and scope, whereas open access analytics focuses on the open access agenda. Hence, data that can serve as a data source for open access analytics should be relevant to open access publishing.

2.2.4 OAA Delivery Using Open Access Repository Infrastructure

One of the early attempts to harness the analytics opportunities offered by open access repository data was not directly relevant to open access analytics. Instead, it was based on the utilisation of the availability and openness of open access literature for research performance evaluation agendas, such as research impact.

Hitchcock et al. (2002) and Brody (2003) introduced Citebase as the first citation database analysis tool built on top of open access repository data. Citebase was designed to perform a set of citation analysis based on bibliographic data extracted from open access literature. The early version of Citebase was solely dependent on arXiv repository data, the subject-based open access repositories. Then, the service coverage was extended by harnessing the availability of standardised protocol to exchange data between open access repositories.

Therefore, a service provider called ‘Celestial’ was developed and operated to support Citebase analysis. Celestial is an OAI-PMH service provider that performs periodic harvesting of a set of repositories indexed by Registry of Open Access Repositories (ROAR). In addition to Citebase, the Celestial service provider was used to operate a ROAR analytics service (Brody, 2006; Carr and Brody, 2007) to measure the repositories’ deposit growth.

In this scenario, the analytics are provided as an added value service on top of the OAI-PMH service provider (Celestial). Open Access Initiative-Protocol for Metadata Harvesting is a low-barrier interoperability protocol based on metadata harvesting that provides a framework for metadata exchange between open access repositories. This framework enables interoperability between two primary entities: the data provider, where the OAI-PMH protocol is implemented and used to expose metadata about their managed resources, and the service provider, which harvests the exposed metadata to construct added value services on top of the repositories' metadata.

However, OAI-PMH service-provider development and operation are problematic processes. Liu et al. (2005) report the case study of the Arc service provider, the first OAI-PMH service provider. They note that the project encountered a set of challenges, including metadata inconsistency, a lack of control vocabulary and Extensible Markup Language (XML) errors. Liu et al. (2002) argue that the OAI-PMH–based applications are associated with quality challenges in data and service availability. While data quality issues are raised by low compliance to OAI-PMH standards, the service availability is influenced by factors out of the control of the service provider. Also, they emphasise the scalability issues raised by the growth in number of repositories, including their collection, which influences the harvesting process and makes harvesting task resources a time-consuming process.

While Liu et al. (2002) propose OAI proxy, replication and caching strategies to reduce these challenges, their proposition is based on a centralised system, which places the burden on the specific community that launches and operates the service provider, which is encountered through the digital library grid that aims to distribute the cost of establishing a service provider. Liu et al. (2005) assert the following about the grid digital library:

> [The digital library grid] propose[s] to distribute the cost of publishing to collection builders (data providers), distribute[s] the cost of harvesting and indexing to existing grid nodes, and only leave[s] the cost of maintaining the federated search service to one institution (service provider), thus making it more sustainable. (Liu et al., 2005, p 601)

Another effort made to reduce the complexity of establishing a service provider is the D-NET Software Toolkit (Manghi et al., 2010). D-NET is a general purpose service-oriented framework designed for the aggregation of infrastructure purposes to reduce the cost of software development. However, the value of a service provider is not in its software, but rather in the value of the collected and curated data within a particular service provider (Knoth and Zdrahal, 2012).

One of the recent and ongoing projects that enables analytics on top of open access repository data is ConnectedRepositories (CORE), which aims to provide centralised access to full-text content in their aggregated database. CORE aggregates the distributed open access repository content, providing added value services on top of others, such as a full-text search engine. In addition to the search engine service, CORE provides three levels of access through an Application Programming Interface (API)s , including analytical access, raw data access and transactional information access (Knoth and Zdrahal, 2012).

However, the challenges of data quality are still present in operating a CORE-model service provider. Thus, a CORE project utilises collaboration within its system by using a dashboard provided for open access repository managers to collaborate in terms of resolving metadata issues and controlling the harvesting process (Pontika et al., 2016).

2.3 The Concept of the Social Machine and Web Observatory

The delivery of open access analytics using open access repository data is a collaborative task between the data provider and the service provider. As discussed above, in its basic form, a distributed repository feeding a centralised aggregator operates its service on top of the harvested metadata. The level of collaboration within the community is enhanced by adopting a crowdsourcing solution pool of the distributed repositories' profiles and providing it from a single point of access. ConnectedRepositories, and its portal, move this collaboration to another level. In ConnectedRepositories, the repositories' management is motivated and powered by the capabilities to engage with the service providers in the harvesting process and resolve metadata issues in their repositories.

On the other hand, the concept of social machines is harnessed to reconstruct a complex task through orchestrating the collective effort of groups of participants. For instance, Murray-Rust et al. (2015a) utilise the social machine model to enable collaborative community-based software development. The development and enhancement in the collaboration of operating open access analytics on top of OAI-PMH infrastructure and the adoption of the social machine model to realise various forms of tasks, provides inspiration on how the social machine concept can be utilised to reform the open access analytics process. However, this aim cannot be carried out without a basic understating of what a social machine is. Thus, this next section discusses the concept of the social machine to articulate the debate on what is a social machine and adopt the working perspective of this doctoral thesis.

Besides the social machine concept, the web-science research community coined and developed the concept and implementation of *Web observatory.* Web observatory is envisioning an open environment for analytics. This environment is incorporated in this research as an intermediate layer of support for open access analytics . A further detailed dissection is provided in Chapter 5. Although, this section provides overview on web observatory concept and architectural.

Understanding the concept of a social machine and the concept and implementations of web observatory is necessary to understand the conceptual framework, analysis and development of open access analytics using open access repository data with a social machine (OAA-OARD-SM) in Chapter 5.

2.3.1 The Concept of the Social Machine

The concept of the social machine is an emerging notion that is regarded in various ways by a number of research communities. One attempt to unify these perspectives can be found in Buregio et al. (2013)'s review. Buregio et al. (2013) categorise the concept into three main ideas: people as computational entities, socialising entities and social software. However, since Buregio et al. (2013)'s review was published, a set of perspectives have been provided that require effort to align with these three ideas. Instead, this work draws attention to those perspectives as two main general views, specifically, the views of web science and software engineering.

2.3.1.1 Web Science Perspectives

Web science researchers consider social machines as new interpretations of existing phenomena, where technology and human aspects are investigated as coexisting components (Tinati and Carr, 2012; Shadbolt et al., 2013). The primary notion of this interpretation is provided by Berners-Lee (2000, p 172-175) in their book Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web, which describes the social machine as "... processes in which the people do the creative work and the machine does the administration ... to create new forms of [a] social process would be given to the world at large, and development would be rapid."

As emphasised by Hendler and Berners-Lee (2010), Berners-Lee (2000)'s view that social machines can be applied to some new paradigms was defined after Berners-Lee (2000) used the concept in Web-related contexts, including crowdsourcing, social networks and collaborative platforms with successful projects, which emerged based on these paradigms. Therefore, Shadbolt et al. (2013) drew a comparative view based on the focus of these related paradigms, as summarised in Table 2.3, to define the boundaries of the concept of social machines compared to other paradigms.

Related paradigm	**Paradigm focus**	**SM envision**
The Wisdom of the Crowd and Collective Intelligence	Wisdom of the Crowd is a decision-making purpose paradigm based on considering a group of people's opinion through the use of technologies. Collective Intelligence focuses on identifying a situation with a group of people instead of an individual	The way computational intelligence and human cooperation unite to achieve a given purpose is more automated, and the relation between them is a co-existing relation.
Open Innovation	Open Innovation is a paradigm that motivates firms to use internal and external ideas to developed internal or external paths to market.	The level of interaction between the social and machine-driven processing components is more in social machines.
Human Computation	Human Computation is an AIcentric paradigm with a purpose to overcome the computer limitations to perform some tasks by applying human processing power.	Not AI focussed and covering wider scenarios where people can be used to perform new tasks not presented by the Human Computation community.
Computer-Supported Collaborative Work and Social Computing	Computer-Supported Collaborative Work is more focused on the information management capabilities of groups and communities, rather than the way these capabilities emerge as a joint effort. Social Computing is the aggregation, presenting, processing, use, and dissemination of information that is distributed across social collectivises including teams, communities, organisations, and markets.	The social and the technical components are equal and necessary partners and study the ways they could be best combined to master the challenges of future socio-technical systems.

Table 2.3: Shadbolt et al. (2013) comparative view of related paradigms

The Berners-Lee (2000) definition is based on the functional contribution of human and machine elements, with the human element presented as a creativity role and the machine element as an administrative role (Smart et al., 2014). Given such constraints, Smart et al. (2014) redefined social machines as

> Web-based socio-technical systems in which the human and technological elements play the role of participant machinery concerning the mechanistic realisation of system-level processes. Smart et al. (2014, p 5)

According to Smart et al. (2014), this definition highlights that social machines are web based, not as an inclusion or exclusion criterion, but to draw the attention of web-science researchers to web-based social machines. Another important aspect of social machines, as highlighted by Smart et al. (2014)'s definition, is their socio-technical characteristics, which are also emphasised by Tinati and Carr (2012). Tinati and Carr (2012) assert that social machines of all scales share a socio-technical structure. They are a network of actors that depend on the relationship between society and technology, constructed collaboratively. Similarly, Shadbolt et al. (2019) stated that:

> The sociotechnical nature of social machines is not hard to grasp: no people => no social machine, and no networked digital technology => no social machine either. (Shadbolt et al., 2019, p 44)

A socio-technical system is a system of social entities interacting with each other with support from technical components for computing and communication (Chopra and Singh, 2016) Furthermore, it is a system that relies on humans and technology to operate, as the technical objects of the system are treated as equifunctional to human capabilities, except in the goal definition of Ropohl (1999).

This leads to the third aspect asserted by Smart et al. (2014)'s definition, which is that a social machine requires the joint involvement of humans and machines in terms of the execution of particular processes. However, Smart et al. (2014) note that the joint involvement and integration of human and technological capabilities occurs more in social machines than in conventional socio-technical systems. Tinati and Carr (2012, p 975) state that "any task that requires the co-constitutional involvement of human and technology is a form of [a] social machine".

However, Smart et al. (2014) argue that the task itself does not mean that it is a social machine. Although they assert that the process is a central aspect of their understanding of social machines, they state that it is the physical system that performs, imports and releases the process.

Within the web-science research community, another perspective is provided by Luczak-Roesch et al. (2016, p 559), who describe social machines with a broader view, stating that they are “the emerging output of human activities rather than any fixed engineered input”. Their theory is based on a principle obtained from their prior work (Luczak-Roesch et al., 2015), which is that the accumulation of information sharing activities on the Web leads to purposeful collective action. This theory describes social machines as a situation created by a set of dynamic actions co-constituted by human and computer elements. Hence, this perspective matches that of Shadbolt et al. (2013) and Smart et al. (2014), that there is a coexistent relationship between human and computer elements in social machines. Therefore, this definition asserts that social machines are a form of output, rather than a system.

Regarding different points of view on what a social machine is, Tarte et al. (2015) introduce the idea that social machines are entities formed of elements that are participants in the social machine, rather than a user including people, algorithms, data objects and infrastructure. Thus, social machines present a major conceptual shift in the conceptualisation of the computational and human elements’ role in the socio-technical system. Tarte et al. (2015, p 25) define social machines as ”entities integrating social energies and computational powers into a sociotechnical system (whether purposeful or not) where social dynamics animate communities”.

In short, they assert the key role of the integration of social capabilities and computational power to release a particular process. This view is in line with Smart et al. (2014)’s perspective on the participatory role of both humans and machines in the realisation of system processes, as well as the importance of the integration of human and technological capabilities within social machines.

2.3.1.2 Software Engineering Perspectives

On the other hand, the software engineer’s vision of a social machine is based on revealing the power of connectivity that enables the socialisation among software entities (web services) on a global scale. This socialisation process establishes a network of dependencies that form an ecosystem defined as a ‘mash-up ecosystem’. Therefore, this vision arose after investigating the structure and dynamics of these ‘mash-up’ ecosystems using network analysis (Yu and Woodard, 2009) or treating their formation as clustered communities (Maamar et al., 2007) based on functionalities, non-functional properties (Benatallah et al., 2003) or domain of interest (Medjahed and Bouguettaya, 2005) to form loosely coupled information systems (Maamar et al., 2005).

In contrast, some software engineer researchers ‘tweak the local’ by redefining the entity with a new mantel model, considering the Web a programmable platform. Hence, a new representation model, taking into account the relationship and interaction with other entities, has been established. For example, Meira et al. (2011) define a social machine as:

> A connectable and programmable entity containing an internal processing unit (P) and a wrapper interface (WI) that waits for requests (Req) from and replies [with responses (Resp)] to other social machines. (Meira et al., 2011, p 26)

This view considers that a social machine is a new entity emerging from the relationship established with other social machines to provide a new service. Therefore, its internal processing unit establishes that "connections define intermittent or permanent relationships (Rel) with other SMs, connections which are established under specific sets of constraints (Const)"(ibid.).

In this case, the relationship is a type of connection that constrains the association and interaction between two or more social machines (Burégio et al., 2013). Taking advantage of the conceptualisation of information-processing systems introduced by Burgin (2006), Burégio et al. (2013) redefine the term social machine to mean

> A connectable and programmable building block that wraps (WI) an information processing system (IPS) and defines a set of required (RS) and provided services (PS), dynamically available under constraints (C) which are determined by, among other things, its relationships (Rel) with others. (Burégio et al., 2013, p 47)

Hence, the internal processing unit is replaced with an information processing system that is comprised of hardware, software and infoware. As this model has been investigated in different contexts, including governmental systems (Burégio et al., 2015), enterprise applications (Burégio et al., 2015), social networks (Burégio et al., 2013) and personal API (Buregio et al., 2014), a new annotation language has been established to describe the network of relations between social machines, known as Social Machine Architecture Description Language (Nascimento et al., 2014).

2.3.1.3 Summary & Adopted Perspective

Given the varying perspectives of social machines, it is essential to adopt a working perspective consistent with the uses of the concept in this report. In terms of the socio-technical aspects in the core principles of web science, there is no doubt about its importance in software design and engineering. Therefore, the absence of specific emphasis on the importance of human interaction in the software engineering perspective limits its scope to computer properties only. This report outlines the research in web science, and therefore, the perspective of web-science researchers is adopted. The definition by Luczak-Roesch et al. (2016) is very general and does not utilise the existing efforts in socio-technical system engineering unlike Smart et al. (2014)'s definition, which draws a definite link between socio-technical systems and social machines.

To sum up, this thesis adopts Smart et al. (2014)'s definition of social machines, which characterises them as follows:

- **Social machines are socio-technical systems**: they involve both human and technological elements in the realisation of the system process.
- **Social machines involve multiple human participants**: a social machine harnesses the capabilities of a large number of participants in the system process.
- **The processes are central to the existence of social machines**: they make sense of social machines when the process is realised by joint contributions of multiple human participants and technological components.

2.3.2 The Web Observatory

Berners-Lee et al. (2006) ccall for the establishment of a new science dedicated to understanding the phenomena raised by web-user interactions. They put forward the view that it is necessary to ensure engineering is a beneficial and powerful new tool and is done 'doing it with our eyes open'. However, an analysis of web activities is a challenging task due to the heterogeneity of the structure and domains, as well as the fast-growing and highly dynamic nature of the web environment. Thus, it is not feasible to depend on a single research group or institution to analyse it (Tiropanis et al., 2013). To overcome these issues, a web observatory is proposed as a new framework to provide the web-science community with the distributed infrastructure to collaborate on the analysis of web activities (Tiropanis et al., 2014c).

2.3.2.1 Definition

The concept of a web observatory corresponds, in its essence, with a former instrumental initiative called 'Virtual Observatory'. The Virtual Observatory was a collaborative project carried out by the International Virtual Observatory Alliance (IVOA) [1] to promote the collaboration of researchers collecting data and analysis on astronomical projects. Later, the web-science community adopted the observatory concept to support their domain and underpin the interdisciplinary nature of web science.

Tiropanis (2012) describes the web observatory as

> A distributed archive of data on the Web and its activity, and, at the same time, the mechanisms and tools that will be used to explore its development in the past, to examine its present condition, and to establish potential developments in the future.

[1] The International Virtual Observatory Alliance (www.ivoa.net) is an organisation that debates and agrees the technical standards that are needed to make the VO possible.

Thus, a web observatory of data and methodologies, constructed in a distributed form, can be used to understand web development. Another perspective is provided by Tinati et al. (2015). Tinati et al. (2015)'s view is that data, and its analytical applications and visualisations, are built by humans as the Web is constructed. Hence, infrastructure consists of data and applications which are opened, linked, owned and controlled by their users (Luczak-Roesch and Tinati, 2017). Tinati et al. (2015) also describe a web observatory as a distributed infrastructure that supports resource sharing, while the privileges to view, query and download are controlled by their owner.

Brown et al. (2014) provide taxonomical facets extracted from web-observatory literature and case studies. According to their taxonomy, a web observatory consists of five main categories of facets: data-related facets, platform-related facets, interface-related facets, service-related facets and actor-related facets. In addition, a general classification of web observatories, based on their purposes, including academic, business, personal and governmental, are provided.

Tiropanis et al. (2014c) distinguish web observatories from other large-scale efforts to analyse and archive the Web. A web observatory is established through distributed and collaborative efforts. Therefore, it is not managed and maintained by a single administrative domain, unlike close cooperative efforts, such as Google Analytics[2]. Also, web observatories should utilise the Web's infrastructure and standards, in contrast to grid infrastructure projects where peer-to-peer agreement is adopted.

2.3.2.2 Architectural Principles

To achieve the vision of web observatories, the research community should ensure that four architectural principles are considered and provide the community with a wrapper to accommodate different scenarios and application domains (Tiropanis et al., 2014a).

- *The architecture allows different types of licences to be managed.* A web observatory is proposed as a solution to bridge the gap between big data and private data, where commercial and military use may be involved. Therefore, access should be controlled (Tiropanis et al., 2014a). Also, a wrapper is proposed where the owner of the dataset can exchange and analyse the data they generate or collect (Gallen, 2013).

- *Both the data and analytical application built on top of the catalogued data should be provided with an explicit link to the data and its analytics.* A web observatory is not just an internet archive where datasets are provided but includes both the

[2] Google Analytics (analytics.google.com) is a web analytics service offered by Google that tracks and reports website traffic, currently as a platform inside the Google Marketing Platform brand.

data and the tools required for analysing and visualising datasets. Its implementation supports all phases of web-content analysis, including general- and specific-purpose information extraction tools, database systems with the ability to handle heterogeneous data structures and technologies and analytical tools supported by visualisation capabilities (Tiropanis et al., 2014a,c).

- *Both local and remote hosted resources can be catalogued.* Due to the large amount of data available on the Web, cooperative work in data collection and analysis is crucial. Therefore, one of the architectural principles of a web observatory is for it to be built as a distributed infrastructure, where tools and datasets are shared within a network of individual web observatories (Tiropanis et al., 2014a)
- *Components and resources globally identified and described with metadata are published to be harvested and used to access web observatory resources.* The datasets are described within the individual system using metadata and published to identify the resources supporting the dataset and tool visibility within system networks. DiFranzo et al. (2014) propose a new extension of the shcama.org project to describe individual web observatory projects and their components using a semantic model extended from a global project.

2.3.2.3 Web Observatory Network

As the data sources, observation methodologies, and analysis approaches are varied, the trend of current and proposed web-observatory project implementations are based on a highly scalable approach, where the web observatory implementations adopt an easy plug-in and plug-out of data sources and analytical tools, such as Southampton University Web Observatory (Tinati et al., 2015) and Next Live Observatory (Luan et al., 2013).

Within individual web observatories, a portal should be presented, which acts as an end point for community engagement. This portal should also be underpinned by a set of technologies that supports its tasks, from source allocation to dataset exchange. For example, Southampton University Web Observatory (SUWO)'s implementation is comprised of three separated main components bridged by API . These components are data stores, analytical applications (visualisation and analytical tools) and a web observatory portal, as shown in Figure 2.4.

This type of architecture enables an individual web observatory to scale horizontally, depending on hosted data and analytical application. Both datasets and analytical tools are presented and described using metadata in the web-observatory portal to allow engagement within web-observatory networks. Thus, the infrastructure can scale to form an interconnected network consisting of a set of web observatories.

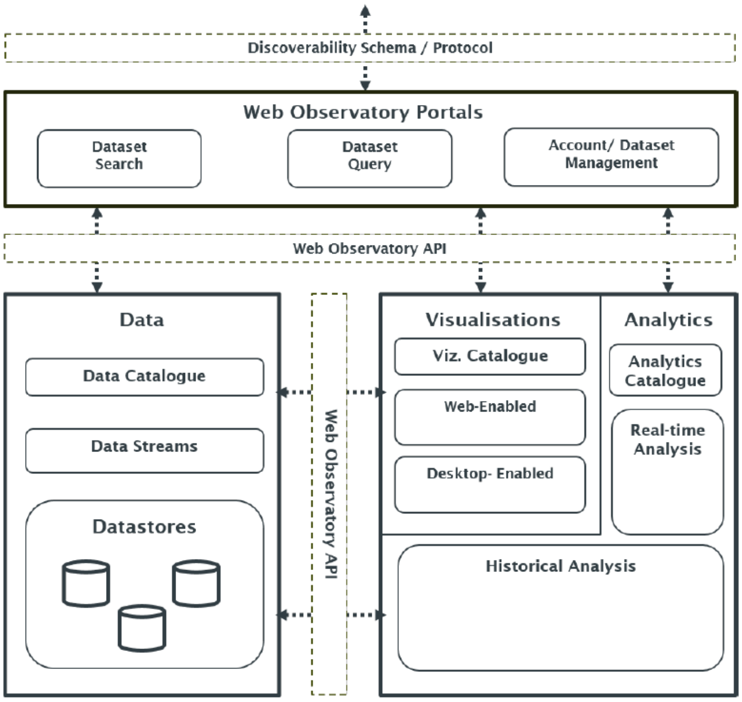


Figure 2.4: University of Southampton Web Observatory (Tinati et al., 2015)

2.4 Chapter Summary

This chapter discusses the concept of open access analytics through a narrative review of open access publishing, data analytics and open access repository literature. In summary, open access analytics is the data analytics process characterised by open access publishing as the analytics object, despite the analytics subject being used in the analytics process. Also, the open access repositories present an advantage to delivering open access analytics on top of their data. Therefore, their position in open access publishing as a strategic route to make research open access and its scope to collect various forms of open access research, was demonstrated.

Furthermore, this chapter provides a fundamental overview of social machines and web observatory concepts. To sum up, the social machine is a web-based socio-technical system in which human and technological elements play the role of participant machinery concerning the mechanistic realisation of system-level processes. However, web observatories are an open environment for analytics, supporting the community with a distributed archive, access control over their data and collaboration in the method development and process of analytics.

In this doctoral thesis, Cooper et al. (2012)'s definition of data analytics is adopted, which defines analytics as the process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/or simulated future data. Cooper et al. (2012)'s perspective on data analytics is one of the perspectives that conceptualises analytics as a process, instead of highlighting its architectural and technical systems. This perspective orchestrates this doctoral thesis' perspective on data analytics and the adopted perspective on the social machine concept that positions the process as a central component in understanding a particular social machine. Accordingly, identifying the process of open access analytics is determined as one of the primary phases of reconceptualising the delivery of open access analytics as a social machine in Chapter 5.

Chapter 3

Research Methodology

3.1 Introduction

The research methodology is a framework of systematic procedures and justification used to solve the research problem (Kothari, 2004).It provides methodological approaches composed of a set of philosophical assumptions, research methods and research design (Creswell, 2009).This chapter offers a theoretical foundation with regard to the research methodology used in this study. In addition to introducing the research approach and methods, it also provides a justification as to *why* such methods were selected and explains *how* the research was carried out. On the other hand, the research procedures are provided with their related studies in the subsequent chapters.

3.2 Multi-Methods Qualitative Approach

Research is a process of inquiry and investigation. This process is governed by basic, broad and overall assumptions that impact the decisions and procedures within a particular study; these overall assumptions constitute the research approaches. In his book *'Research design : qualitative, quantitative, and mixed methods approaches'*, Creswell (2009, pp 3) defines research approaches as: "... the plans and the procedures for research that span the steps from broad assumptions to detailed methods of data collection, analysis, and interpretation".

The term 'approach' is used interchangeably with research strategy or research philosophy. For instance, Amaratunga et al. (2002) use the term when differentiating between positivist and interpretivist philosophies in social research, while Crowe et al. (2011) and Mabry (2008) describe the research strategy of case studies as an approach. However, Creswell (2009) highlighted both the research worldview and research strategy, in addition to the research methods, as components of the research approach. In light of this, he lists three main approaches to research: quantitative, qualitative and mixed methods research.

Qualitative and quantitative research approaches each collect one form of data and adopts one methodological tradition. The quantitative research approach is the dominant objectivist approach, and is based on testing theories deductively by raising hypotheses and confirming them through the collection of quantitative data and the use of analysis methods associated with a rigorous, appropriate sampling strategy (Robson and McCartan, 2016). While the generalisation of findings is essential in quantitative research, this is not the case when using a qualitative approach (Maxwell, 2013). Indeed, qualitative research is conducted in order to seek a better understanding of phenomena through in-depth investigation, using qualitative data collection techniques such as interviews, open-ended questions or observations (Creswell, 2009).

Mixed-methods research utilises both quantitative and qualitative approaches conducted in parallel or sequentially, with the output integrated and triangulated to gain a better understanding of the research problem at hand (Creswell, 2009). Although the mixed-methods approach is recognised as an independent approach, qualitative and quantitative approaches serve as central to its existence. This is due to the fact that it utilises the strength of each approach to obtain more credible findings, given that all methods and approaches have their advantages and disadvantages. Consequently, a combination of methods can strengthen a study and lead to more reliable results (Doyle et al., 2016).

Another approach that adopts a similar assumption is the multi-method approach, which has been recognised as a broad umbrella term, including both multi-method and mixed-methods research. Despite the fact that the terms ‘multi-method’ and ‘mixed-methods’ are used interchangeably, there are researchers who attempt to draw a definite line between the two. For example, while the integration of qualitative methods and quantitative methods is identified as an essential methodological component in the mixed-methods approach, this is not the case with the multi-method approach (Anguera et al., 2018).

The multi-method approach is the combination of different styles of research within a single research project, although is not restricted to mixing both qualitative and quantitative methods. This is in contrast to the mixed-methods approach, which involves the combination and integration of both quantitative and qualitative research methods (Johnson et al., 2007). To clarify, Morse (2003) describes the multi-method approach as a form of research involving multiple types of qualitative or multiple types of quantitative methods. That is to say, the methods come from the same tradition, and the overall research is underpinned by one philosophical worldview that supports the adopted tradition (Creswell, 2009).

3.2.1 Qualitative approach with Interpretivism

This research uses a qualitative multi-method approach to acquire a greater understanding of conducting open access analytics using open access repository data. In addition, a sub-set of research questions were derived based on the sub-aims and objectives, in order to support the overall aim. This qualitative multi-method approach takes the interpretivist theoretical perspective (Robson and McCartan, 2016).

Indeed, qualitative research has such a strong link to interpretivist philosophical assumptions, that it has been considered equal to interpretivism (Goldkuhl, 2012). Despite the debate on this extreme interpretation of the correlation between qualitative and interpretivist approaches, interpretivist philosophical assumptions are in harmony with the nature of qualitative research, which is based on inductive elicitation of understanding from qualitative non-ordinal data (ibid.).

Two of the core philosophical assumptions that constitute the interpretivist worldview are constructivist ontology and constructivist epistemology (Robson and McCartan, 2016). Ontology refers to what is, the nature of existence and what constitutes reality (ibid.). However, constructivist ontology is built on the assumption that the world is not 'given', but is instead 'produced and reinforced by humans through action and interaction' (Orlikowski and Baroudi, 1991, pp 14). Epistemology is about what mean to know and what constitutes human knowledge (Robson and McCartan, 2016). Constructivist epistemology is based on the belief that truth and meaning are created by human interactions with the world (ibid.). Thus, it is not discovered, but rather constructed, and subjects construct their own meaning when it comes to the phenomena under investigation. Therefore, the role of the researcher is to uncover these meanings as a valid account of reality and the phenomenon in question (Goldkuhl, 2012).

Based on these broad assumptions, an interpretivist adopts an approach that results in understanding through interpretations of the problem or phenomenon at hand. The interpretivist researcher interprets the 'existing meaning systems shared by the actors' (Orlikowski and Baroudi, 1991, pp 15), therefore deriving a holistic understanding through the use of the hermeneutic circle and contextualisation (Goldkuhl, 2012). While contextualisation is based on the concept that the acquired knowledge is situation dependent and needs to be understood based on the context of the phenomenon (Madill et al., 2000), the hermeneutic circle is the process of moving between the whole and the parts of the phenomenon to acquire a holistic understanding (Klein and Myers, 1999).In this case, the researcher is an instrument embedded in the social context, using their skills and their relationships with the participants to gather information about the phenomenon, and using their knowledge and insight (Bhattacherjee, 2012), as well as the existing knowledge and literature, as a 'synthesis device' to provide a clear view of the phenomenon at hand (Klein and Myers, 1999, pp 75). There is no right or wrong way of doing so, rather, “there are interesting and less interesting ways to view the world” (Walsham, 1993, pp 6).

3.2.2 Multi-Method Research Design

Multi-method research is characterised by the coexistence of various methodologies, such as the combination of observation and ethnographic methodology or the combination of a case study with a survey (Anguera et al., 2014, cited in Anguera et al. 2018). Each research method offers a promising solution for another method’s problems. However, this does not imply that the multiple methods are directed towards answering a single research question or achieving a single goal. Instead, the multi-method approach involves using multiple methods to answer multiple research questions with multiple aims within a single study (Anguera et al., 2018).

In addition, the multi-method approach can contribute to a single, broader research question or support the investigation of a research problem. Hesse-Biber and Johnson (2015) describe the multi-method approach as a situation in which a particular study calls on a second study using a different method from the same tradition, and which thus plays a secondary role to the primary study. Anguera et al. (2018) emphasised the importance of each method going through four main steps of scientific inquiry: identification and formation of the research question, data collection, analysis, and interpretation of findings. The overall research design is illustrated in Figure 3.1.

This research uses a narrative literature review (see Section 3.3.1) to provide the reader with the background on open access analytics as a concept synthesised from data analytics and open access publishing literature. It then scales down the focus of the research solely to the open access analytics that uses open access repository data. research problem concerns the approach used to process the open access repository data for open access analytics. In addition, a background is also provided regarding the concept of social machines and web observatories, so as to support the research inquiry in RQ2 and the interpretation of the research findings related to RQ3.

The rationale around incorporating a narrative literature review is based on the purpose of the literature review in this study. Research into data analytics practices within the open access repository literature is scarce and provided as technological systems and the architecture of services providers (Knoth and Zdrahal, 2012; Lossau et al., 2006; Robinson and Horstmann, 2007; Rettberg and Schmidt, 2012). This can be attributed to the fact that open access repositories are acknowledged for their role as scholarly communication forums (Lynch, 2003) and digital preservation entities (RLG-OCLC, 2002; Allinson, 2006; Nicholson and Dobreva, 2009), rather than being recognised for supporting analytics practices. Therefore, this research takes advantage of the data analytics literature to provide an understanding its application in open access repositories. Furthermore, the literature on open access repositories, open access publishing and data analytics is incorporated to provide a narrative literature review on the subject of open access analytics.

This thesis begins with an examination of ROAR analytics as a case study, focusing on the process of deriving the analytics service on top of open access repository infrastructure, and investigating the OAI-PMH service provider as an approach to processing the repositories data. According to Yin (2009), using a case study is a strategy that can incorporate both quantitative and qualitative data collection; however, overall, case studies tend to be qualitative in their design. In addition, case studies can take an exploratory, descriptive or explanatory form. To answer RQ1, an explanatory case study was conducted. The challenges and limitations of a service provider's approach using an OAI-PMH interoperability layer are discussed in the literature (Liu et al., 2002, 2005; Houssos et al., 2011; Meschenmoser et al., 2016). However, these studies are not associated with data analytics and open access analytics requirements, and thus an explanatory case study can provide a better understanding of the causes and effects (Yin, 2011) of analytics application on open access repositories using an OAI-PMH service provider. It can also resolve the challenges associated with the dynamics of analytics application requirements. Furthermore, this thesis harnesses the value of one of the cases significantly depends on the OAI-PMH to generate its analytical assets. In addition, it takes advantage of the opportunities offered by the OAI-PMH, which can enable a greater understanding of the support provided by the OAI-PMH framework for open access analytics practices.

What follows is an exploration of the concept of social machines that are contributed by the web science research community, which presents a new form of joint collaborations of human participants and machines through the web-based social technical system (Smart et al., 2014). The second aim of this study is to incorporate the process of social machines into open access repositories infrastructure to carry out open access analytics. The social machine research community values the process within a particular social machine to understand or identify the social machine (ibid.). Thus, the process of open access analytics using an open access repository is sufficient to achieve the second aim. While the ROAR case study presents the analytics process in light of ROAR analytics requirements, the open access analytics process cannot be generalised or deduced from a single case. Nevertheless, in the data analytics research domain, there are a number of research communities that have drawn attention to the analytics process and conceptualised it with a number of process models. Therefore, these conceptual models are used to conceptualise the open access analytics process in the open access repository literature, through constructive conceptual analysis (see upcoming Section 3.3.4).

It is important to mention that this part of the study does not aim to engineer new infrastructure or propose a new system; its objective is to synthesise the open access analytics as a social machine process that takes advantage of the existing infrastructure, interactions and communities formed around open access repositories. Therefore, the output of the conceptual analysis is incorporated into this synthesis, in addition to an examination of open access repository stakeholders and their roles, web observatory infrastructure, and the concepts constraining or supporting analytics activities using open access repositories. The output of conceptual synthesis is provided as the Open Access Analytics with Open Access Repository Data with Social Machine (OAA-OARD-SM) conceptual framework.

The social machine research community adopts Lightweight Social Calculus (LSC) to represent a particular social machine interaction (Murray-Rust et al., 2014). This method provides the researcher with interaction simulation techniques on a large scale. Although the LSC provides an approach through which to design and evaluate open access analytics as a social machine process, the purpose and the theoretical perspective adopted in this thesis call for greater understanding and contextualisation of its core principles. Therefore, the conceptual framework is used as 'a synthesis device', to illustrate the process from a broader, conceptual level, given that the aim is to highlight the propositions that require more attention and investigation.

One of the main propositions of the OAA-OARD-SM conceptual framework is the participatory role of open access repository management in the analytics process. Therefore, the third aim of this research is to understand the existing analytics practices within the boundary of open access repository management. Expert interviews were incorporated in order to answer the third research question. These expert interviews enabled the researcher to gain valuable knowledge from the experts, by exploring their practices and experiences of the issue being studied. In the expert interviews, the repository management team members were used as experts who reflected subject matter expertise (see upcoming Section 3.3.6). Data regarding their level of experience and their position in relation to the open access repository was collected using a short questionnaire (see appendix A), and those with expertise were then interviewed. A semi-structured interview was used to collect rich data guided by a set of agenda (see Appendix B). This rich data was then analysed using thematic analysis, in order to identify the applications of the open access repository data for analytics practices, the interactions of the repository managers within these applications and the concerns of the recruited experts regarding these applications. The findings from the expert interviews were used to revisit the OAA-OARD-SM and refine it, taking advantage of the new understanding, acquired from the interviews, of open access analytics at an open access repository level.

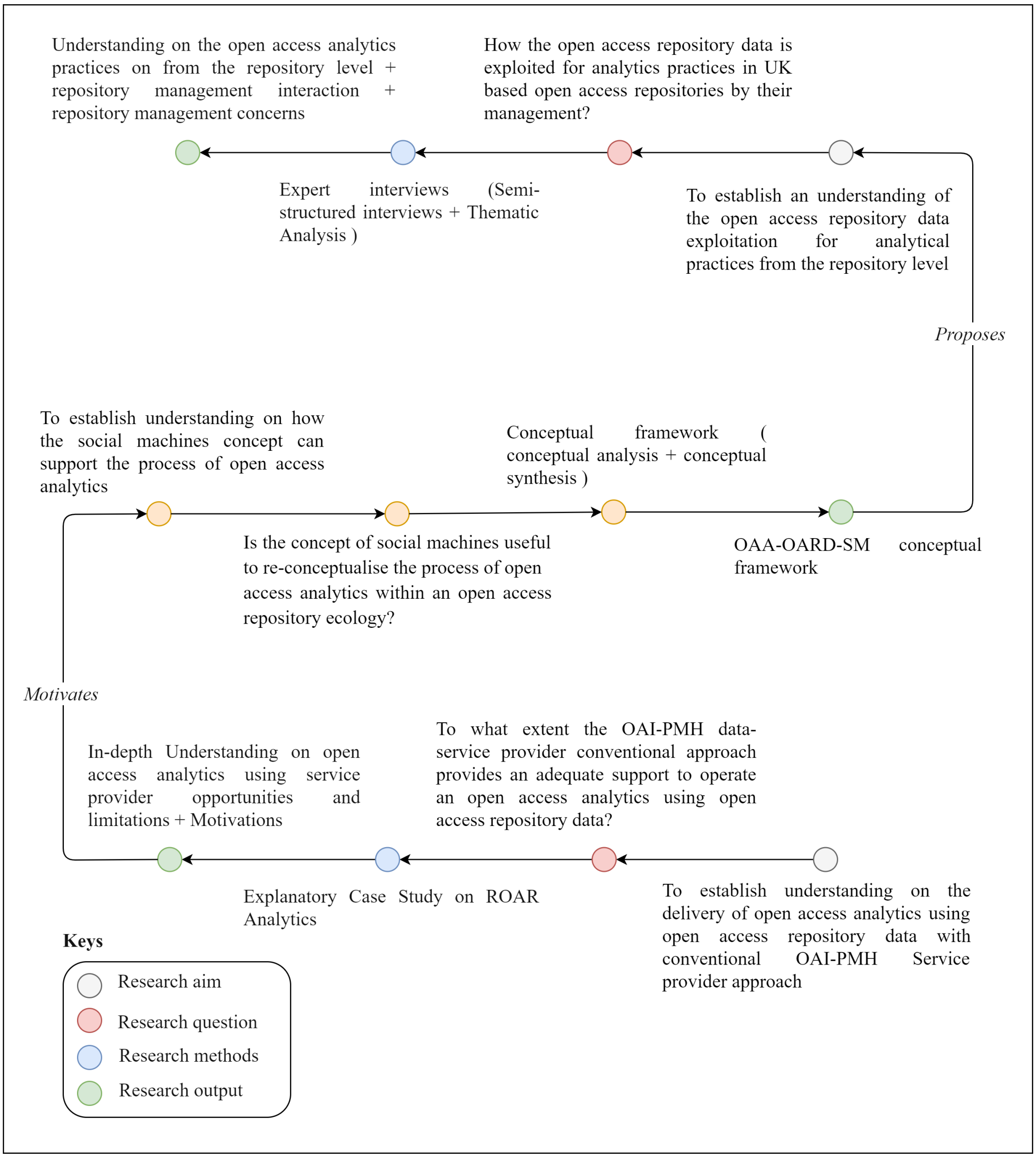

Understanding on the open access analytics practices on from the repository level + repository management interaction + repository management concerns
How the open access repository data is exploited for analytics practices in UK based open access repositories by their management?
Expert interviews (Semi-structured interviews + Thematic Analysis)
To establish an understanding of the open access repository data exploitation for analytical practices from the repository level
Proposes
To establish understanding on how the social machines concept can support the process of open access analytics
Conceptual framework (conceptual analysis + conceptual synthesis)
Is the concept of social machines useful to re-conceptualise the process of open access analytics within an open access repository ecology?
OAA-OARD-SM conceptual framework
Motivates
In-depth Understanding on open access analytics using service provider opportunities and limitations + Motivations
To what extent the OAI-PMH data-service provider conventional approach provides an adequate support to operate an open access analytics using open access repository data?
Explanatory Case Study on ROAR Analytics
To establish understanding on the delivery of open access analytics using open access repository data with conventional OAI-PMH Service provider approach
Keys
Research aim
Research question
Research methods
Research output

FIGURE 3.1: The overall research design used in the thesis

3.3 Methods

According to Bielik et al. (2014, cited in Kosterec 2016), a research method is a set of instructions that fulfils a specific goal. This section briefly introduces the research methods and techniques used in this study.

3.3.1 Narrative Literature Review

The literature review is a critical analysis of the relevant research and non-research materials about research problem being investigated (Hart, 1999). It provides a summary, builds greater understanding, keeps the reader up-to-date with the relevant literature and provides a justification and motivation for the investigation of the problem in question (Cronin et al., 2008). There are two main types of literature review: the narrative literature review (also called a traditional literature review), and the systematic literature review (ibid.). In addition, a third, less common type is the integrative literature review (Torraco, 2005).

While a narrative literature review depends on selected material to be a source of knowledge, and does not, therefore, always provide the relevant literature systematically for the reader, the systematic review follows a rigorous, well-defined and systematic approach, reviewing the literature relevant to a specific research topic. This critical difference between them is the manner in which they are carried out, which in turn depends on their purpose. For example, the narrative literature review aims to provide a comprehensive background that informs the reader of the current knowledge and highlights the significance of the research. In contrast, a systematic review prioritises the list, as it examines a complete list of relevant works with a rigorous identification, evaluation and synthesis process (Cronin et al., 2008). Similar to narrative and systematic reviews, the integrative literature review aims to synthesise and critique the literature representative of a particular topic; however, it is distinguished by its output. While narrative and systematic reviews provide a summary understanding of the research topic, the integrative literature review generates new knowledge, new ideas or a new framework relating to the research topic, by either addressing two or established new topics (Torraco, 2005).

3.3.2 Explanatory Case Study

Case study research is based on a detailed examination of an individual aspect of a phenomenon or problem. According to Abercrombie et al. (2006), a case study is:

> The detailed examination of a single example of a class of phenomena, a case study cannot provide reliable information about the broader class, but it may be useful in the preliminary stages of an investigation, since it provides hypotheses, which may be tested systematically with a larger number of cases. (Abercrombie et al., 2006, pp 34)

Therefore, the case study approach provides the researcher with a strategy by which to conduct an in-depth, preliminary investigation of a single case, which can lead to the formulation of a hypothesis. While it is true that hypotheses can be generated from case studies, the notion of limiting the case study strategy to hypothesis formulation, and therefore not assigning a scientific value to it, is problematic. Flyvbjerg (2006) argues that Abercrombie et al. (2006) definition is misleading and limits the case study when it comes to being a primary strategy in research design. Similarly, Yin (2009) considers this confusion a common misconception, based on the belief that the research approach should be organised in a hierarchical manner, with case studies limited to the exploratory phase of the research.

3.3.3 Conceptual Framework

A conceptual framework is used across many scientific disciplines to scope the research out from the conceptual level of abstraction. Miles et al. (2013) define a conceptual framework as a written or visual presentation that:

> explains either graphically, or in narrative form, the main things to be studied - the key factors, concepts or variables and the presumed relationships among them.(Miles et al., 2013, pp 20)

On the other hand, Jabareen (2009) argues that a conceptual framework should consist only of concepts. Thus, he defines it as a 'network or a plan of interlinked concepts' that, as a whole, form a comprehensive understanding of a particular phenomenon. Thus, a conceptual framework is a collection of concepts, in which each of these concepts plays an integral role. Furthermore, the purpose of this construct is to enable a better understanding of a particular topic and provide an interpretative approach, rather than providing a theoretical explanation or analytical setting for a particular phenomenon. Therefore, in some areas of research, the conceptual framework helps clarify the problem's definition, assists with the formation of research questions and provides the justification and motivation for conducting a particular study (Maxwell, 2013).

In a conceptual framework, the set of concepts and the presumed relationships between them are deduced based on the source of knowledge, as well as the researcher's perspective about the area being studied. According to Maxwell (2013), a conceptual framework can constitute the researcher's experiential knowledge, existing theory and research, pilot and exploratory research, and thought experiments.

In addition to the various types of sources, the source of knowledge can be multidisciplinary in nature, that is to say, the role of the conceptual framework is to scope the subject out being studied from a multidisciplinary perspective (Jabareen, 2009). In other words, the role of the conceptual framework is to integrate the concepts from a multidisciplinary source of knowledge, in order to provide an understanding of the phenomenon or define the problem or solutions being studied or proposed (Maxwell, 2013). This is to cope with the absence of a 'skeletal framework' that can be utilised as a source of information within interdisciplinary research (Jabareen, 2009). A skeletal framework is a construct identified by previous research that provides an internal structural guide to the new research process and provides the researcher with capabilities to build on it (Morse et al., 2002).

The use of the conceptual framework concept can vary from one discipline to another. For example, computer science researchers utilise it to describe the abstraction of solution design and any related problems (Mattsson and Bosch, 1997). That is to say, a conceptual framework is a broad abstraction of a particular design that can be reused for a family of similar problems.

3.3.4 Conceptual Analysis

Conceptual analysis refers to a set of instructions that enables the researcher to examine the attributes and characteristics of a particular concept through a formal linguistics exercise (Walker et al., 2005). According to Jackson (1998), one of the recent philosophers who contributed to and advocated the philosophy of concept and conceptual analysis, the conceptual analysis is significant as an integral component of a set of intellectual activities related to the phenomena, and involves categorisation, interpretation and communication. Kosterec (2016), identifies three methods of conceptual analysis, pertaining to its purposes: constructive analysis, detection analysis and reductive analysis.

For each of these three methods, the purpose of conceptual analysis is explicitly revealed in their name. Constructive analysis aims to broaden and extend a particular concept by highlighting its new relationship with other concepts or by refining the existing ambiguous concepts and relations. Detection analysis, on the other hand, does not presume any new relationships, but rather questions the existing relations and may lead to constructive analysis. One of the features of conceptual analysis is its role in ontological reduction; it is questioned whether a particular concept is part of another concept. This form of conceptual analysis is carried out using reductive analysis (Kosterec, 2016).

3.3.5 Conceptual Synthesis

The second conceptual method used in this thesis is conceptual synthesis, which aims to create a new idea or concept from observation, evidence and existing literature. Jabareen (2009) positions conceptual synthesis as one of the phases used to generate a new conceptual framework based on ground ed theory. Furthermore, conceptual synthesis is distinguished from conceptual analysis, as it generates new concepts or ideas based on existing concepts, evidence and literature, whereas conceptual analysis analyses existing concepts (Walker et al., 2005).

3.3.6 Expert Interviews

Conducting expert interviews is a research method that is designed to elicit expert knowledge regarding a particular topic or domain (Bogner and Menz, 2009). Expert interviews can be used to achieve various objectives, which may be contradictory in nature. For example, expert interviews are promoted in empirical social research as an exploratory tool to establish the initial direction of a new or poorly defined research field (Meuser and Nagel, 2009). On the other hand, Bogner and Menz (2009) demonstrate the use of expert review for theory-generation, where the expert interview suited in interpretive philosophical assumption (Meuser and Nagel, 1991, cited in Bogner and Menz 2009), which enables a theory to be generated from a qualitative inquiry. In addition to these two purposes, a third purpose is identified by Gläser and Laudel (2010, cited in Bogner and Menz 2009), where the expert interview is conducted so as to acquire exclusive knowledge possessed by the experts in the field. This exclusive knowledge is action and experience derived from practices in the subject matter, and hence, the expert acts as a guide to the specialised knowledge not otherwise available to the researcher. This form of expert utilisation is described by Bogner and Menz (2009) as a 'systematising expert interview'.

This typological variation in forms of expert utilisation raises various ontological questions regarding who can be described as an expert. According to Froschauer and Lueger (2009, pp 220), experts can be those who are 'equipped with explicit specialist knowledge gained through specific training, which provides them with an in-depth understanding of a particular topic or field'. This in-depth understanding is the expertise that enables them to clarify or resolve specific issues or problems. Similarly, Gläser and Laudel (2009, pp 117) define experts as 'people who possess special knowledge of a social phenomenon'. Both perspectives are based on the view of the expert as a person with substantial knowledge gained through experience, practice and specialised training. Another perspective identifies an expert as a person who has an 'expert role in the investigated social setting' (ibid.), as opposed to being a person equipped with specialised knowledge. Based on this perspective, Gläser and Laudel (2009) define three roles that can be played by an expert in a particular situation: I) Taking the role in the investigated phenomenon as a participant, who may not necessarily be an expert in the field of the study, but rather is part of the phenomenon, and is thus used as an information source about the subject being studied, to reconstruct the phenomenon. II) Taking the role of an expert as a reflection of a particular phenomenon. Even though they not treated as a source of information on the phenomenon, they are an expert in the field, utilised to provide reflections on the phenomenon. III) They are an expert in the field and play a role in the phenomenon, being treated as a source of information for the investigated phenomenon and as an expert in the field.

Froschauer and Lueger (2009) highlighted the forms of expertise possessed when the expert is a person who has had one or more of those experiences . They recognise three forms of expertise:

- **Subject-matter expertise**: A type of expertise gained from practical experience in the field of research, and therefore competent in the procedural knowledge of the field of research.

- **Reflective subject matter expertise**: A form of expertise associated with primary and secondary experience, which is gained from observation and reflection of field knowledge, and consolidates various perspectives about the field of expertise.
- **External expertise**: A secondary experience enabled by relevant theoretical knowledge and second-order observation. It is characterised by a lack of practical knowledge and represents an abstract knowledge about the domain of expertise.

Thus, the type of investigation and the role of any expert knowledge determines the experts that should be recruited, and the methods adopted to elicit the expert knowledge itself. This study adopted expert interviews to answer RQ3 and achieve the third aim of the research, namely to seek an understanding of the analytical exploitation of open access repositories data. This understanding can be achieved by identifying the forms of analytical exploitation and establishing how the repositories data are exploited in particular institutional repositories. Therefore, the expert interviews were conducted in order to elicit both subject-matter expertise (to determine how the data are exploited) and reflective subject matter expertise (to understand the active role of open access repositories data on analytical activities).

In addition to introducing the expert interview as a research method, this section attempts to theoretically justify the procedures of expert selection and recruitment described in Sections 6.2.1 and 6.2.4. According to Pfadenhauer (2009), the concept of ‘competence of experts’ can be used as a means to identify experts for particular research. This involves the extent of privileged access to information and the responsibility for problem-solving decisions. Therefore, the recruitment process for experts on analytical exploitation of open access repositories data should take into account their access to information concerning the internal action and setting of a particular repository, and the individual’s role in the institutional repository in general, as well as in the exploration of open access repository data for analytics purposes.

Accordingly, the repository management have the ability to play the role of experts in the field of open access repository data and to be a source of information on what forms of exploitation exist in the open access repositories. This is due to their primary role in managing the policy, technology and processes of open access repositories, which provides them with substantial access to information and gives them responsibility for the repositories’ management. However, the recruitment process requires extensive experience about the subject matter (institutional repositories) as a prerequisite (Littig, 2009) to ensure the targeting of individuals with substantial expertise concerning analytical exploitation. Therefore, recommendations of well-known experts were utilised in order to purposefully select UK-based institutional repositories with rich information about the analytical exploration of open access repositories data.

3.3.7 Semi-structure Interview

Qualitative research consists of a set of methods and techniques including, but not limited to, interviews, document analysis and thematic analysis. Among these methods, interviews are the most common form of data collection in qualitative research (Legard et al., 2003). Interviews can be unstructured, semi-structured or structured (Gill et al., 2008). The first is challenging to conduct and control, and places a burden on the researcher during the analysis phase, which can lead to inefficiency. However, the third constrains the researcher when it comes to drilling down and exploring the participants' answers further or asking for clarification during the interview. For these reasons, the semi-structured format is often selected, as it provides a balance between the other two forms; although the agenda is defined, the researcher still has the power to intervene, drill down and explore the respondents' answers (ibid.). Such a format was used in this study, enabling rich qualitative data collection from the expert interviews.

3.3.8 Thematic Analysis

Thematic analysis is a qualitative data analysis method used to incorporate a set of systematic procedures in order to identify, organise and offer insight into the meaning (themes) across a qualitative dataset (Braun and Clarke, 2012). One of the important concepts of using thematic analysis is the conceptualisation of the term 'theme'. According to Glisczinski (2018), themes can be conceptualised in thematic analysis as domain summaries or as having a shared meaning-base. While the domain summaries perspective is based on capturing the explicit patterns related to the topic or the issue under investigation, raised directly from the data, the spread of meaning perspective reflects a pattern of shared meaning that can be explicitly or implicitly provided by the qualitative data. In this research, thematic analysis was used with the themes to provide a domain summary perspective.

3.4 Chapter Summary

In this chapter the methodological philosophical standpoint and research design used in this study were discussed. In summary, this research used a multi-method qualitative approach based on the interpretivist theoretical perspective and constructivist epistemological standpoint. The multi-method qualitative approach in this study incorporates a case study, a conceptual framework and expert interviews. These three methods were used to define the procedures that were used during the research process, including the explanatory case study, conceptual analysis, conceptual synthesis, semi-structured interviews and thematic analysis.

Chapter 4

Open Access Analytics using OAI-PMH Service Providers: ROAR Analytics

4.1 Introduction

The discussion on the delivery of open access analytics on top of open access repository infrastructure conducted in Section 2.2.4 shows the cumulative effort of utilising and investigating the data-service provider model using the OAI-PMH protocol as a means of operating analytics functionalities. This reflects the opportunities offered by OAI-PMH for analytics practices. Despite the advantages and opportunities provided by the aforementioned use cases, a set of challenges and limitations exist. However, these opportunities, limitations, and challenges are provided in narrative format to report the cases of OAI-PMH protocol usage and mostly not specific to analytics practices. Therefore, this study attempts to provide a concrete understanding of the degree to which this model can support analytics purposes. This study is driven by the following research question:

> RQ(1). To what extent does the OAI-PMH service provider conventional approach provide adequate support to operate open access analytics using open access repository data?

To clarify, the aim is to explain the causes and effects of the influence of the OAI-PMH as a means to operate analytics and response to analytics requirements, and clarify how open access analytics can be implemented using a conventional OAI-PMH data-service providers approach.

4.2 Case study design

Open access analytics is a process mainly involving problem definition and data analysis to generate analytical insights. Thus, each analytics task is associated with problem definition, goal determination, and objectives that provide a set of requirements that determine the analytics process. Visualisation and data mining technologies enable the analytics process and underpin the insight gathering process, which can be provided as an ongoing service. ROAR analytics are one of the services that provide visualisations regarding the growth of open access repository content and enable insight gathering in terms of the growth and adoption of green open access. The OAI-PMH data-service provider approach is mainly designed for the discovery of repository content, and it is designed with minimal requirements to support its adoption (Lagoze and Van de Sompel, 2003). These analytics activities may have different requirements because the analytics process may require richer data and intensive processing. However, these requirements are dynamic and highly dependent on the analytics problem encountered by the analytics process. To avoid such variation of the analytics requirements, an explanatory case study method (see Section 3.3.2) is used to investigate how analytics can be operated on top of open access repository data and obtain a greater understanding of the degree of support that can be achieved through the OAI-PMH data-service provider approach.

4.2.1 ROAR analytics overview

The distribution of open access repositories without a single point of discovery prevents the open access community from utilising the infrastructure as a single virtual catalogue for the open access literature. Therefore, the open access repository community has established registries to de-fragment the distributed infrastructure and make it work as a service discovery directory. A well-known example of these registries includes a registry of OAI-PMH data providers powered by the Open Archives Initiative (Eprints Group, n.d), the openDOAR (Centre for Research Communications, 2013) project carried out by the Centre for Research Communications at the University of Nottingham, and the Registry of Open Access Repositories (ROAR) (Eprints Group, 2004) powered by the Eprints Group at the University of Southampton.

A similar issue can be found in analytics services established for the purpose of monitoring the adoption of open access repositories worldwide. An example of analytics service promoting and monitoring the adoption of open access repositories is the ROAR analytics service. ROAR promotes open access publishing through a continuous reflection of open access repository deposit activities. This service takes place in single open web-based portals associated with histograms that illustrate and quantify the dynamics of deposit activities to fulfil two main requirements.

RQ1 Repository-Level Quantification: The quantification and identification of open access repositories nodes

RQ2 Item-Level Quantification: The quantification of content in open access repositories to understand deposit behaviours and the growth of this content

The ROAR analytics problem is a part of the ROAR aim, which is to promote the adoption of open access repositories and the growth of their content, as emphasised in their following statement (the goal-related phrases are highlighted):

> The aim of ROAR is **to promote the development of open access** by **providing timely information** about the **growth and status of repositories** throughout the world.

This statement outlines the main aim of the service, which is the **promotion of open access development**. The service investigates the concept of open access adoption on a worldwide scale. Thus, the open access concept should be considered in the data sources (open access repositories). Additionally, it is important to note that the statement defines ROAR's target: focusing on green open access by targeting open access repositories. However, the content of open access repositories is not limited to research output, which is the target of the open access movement as discussed in Section 2.1.1.4, because their content is a combination of learning resources, eprint materials, and research data. Therefore, the datasets used to deliver the service should be relevant to open access development. The statement emphasises the three main characteristics of the service:

C1 **Timely Information** : The service should provide concrete updated information about open access development instead of a one-time analysis or an estimation of this development based on simulated data. Thus, the service should periodically analyse open access development.

C2 **Growth Reflation**: ROAR aims to reflect this development through a growth in open access deposit activities. The growth is cumulatively calculated through a histogram associated with time series data.

C3 **Status Reflation**: A status reflation requires the extrapolation of repositories' status based on growth analysis. Thus, the service assigns a status to a particular repository based on the dynamics of the deposit activities.

4.2.2 ROAR analytics as a case study on open access analytics

The open access repositories deposit quantification service is not a ROAR-exclusive service in the open access community. Like the ROAR service, openDOAR also calculates repository content, but what distinguishes the ROAR deposit monitoring service is the fact that the approach is used to deliver an analytics service. Whereas the openDOAR service is based on systems manually fed by deposit indicators (Millington, 2007), ROAR automatically calculates them using an aggregation system that periodically harvests the OAI-PMH exposed metadata.

This enables a richer form of analysis and provides concrete indicators of deposit activity dynamics. Carr and Brody (2007) have utilised the ROAR monitoring service to recognise patterns in repository deposit activities. These patterns are openly provided for the open access community, visualised into histograms reflecting deposit activities, and aggregated in a single web-based portal. This enables them to introduce a classification schema for the deposit active days, based on the number of deposits per day, and consequently classify the repositories based on the repositories' deposit dynamics. The authors argued that this classification provides insight into the sustainability of repositories.

Fully operating on top of open access repository data and automatically processing the open access repository by taking advantage of the OAI-PMH framework is the first reason for selecting ROAR analytics for the case study. It is correlated to the research problem encountered in this study. The service was established based on another system developed by Tim Brody (Brody, 2006), called 'Celestial', which aims to harvest OAI-PMH metadata from a single source instead of a distributed infrastructure. However, in 2013, Celestial encountered a service failure and a loss of data that led to the suspension of the ROAR analytics service. ROAR analytics represents a contemporary case of analytics with an open access analytics agenda. Further, it faced one of the greatest challenges associated with being a centralised service provider model, which ultimately led to a suspension of its service for several years.

It is important to note that the nature of ROAR analytics and its requirement is typical case requirements. ROAR analytics does not represent a very complex case of analytics, even though it has a very significant role. It merely depends on the data collected from a particular repository, despite the fact that it is implemented on a large scale and provides valuable insights into green open access as a whole. Therefore, the aim of this research is to examine the opportunities of OAI-PMH and its challenges as a typical case instead of an extreme case (Gray, 2014).

4.2.3 Explanatory Single Holistic Case Study Design

There are four basic designs for case study research (Yin, 2009): I) holistic single case design, II) holistic multiple case design, III) embedded single case design, and IV) embedded multiple case design. The decision as to which design is suitable for a particular study is based on the unit of analysis used to answer the research question and the number of case studies intended to be investigated. This study utilises a single holistic design to investigate ROAR data analytics, aiming to uncover the systems, processes, and challenges associated with operating analytics on top of the open access repository infrastructure using the OAI-PMH data-service provider approach.

The determination of which case study design to use for a particular research question correlates with the type of unit of analysis targeted and the number of cases that need to be studied. The unit of analysis represents the type of case the research is investigating, that is, whether it is a person, organisation, event, or problem. The unit of analysis in this study is ROAR analytics as open access analytics, which involves using open access repositories data for open access analytics practices. Whereas the unit of analysis is associated with requirements, processes, and contexts, the design is associated with the utilisation of inside knowledge and log analysis.

4.2.3.1 Inside knowledge

This study investigates one of the services hosted in the University of Southampton, which is the same university the researcher is enrolled in. At the university, the researcher is initiating his doctoral program by recovering the ROAR analytics service through tool development and the re-harvesting of repository metadata. This position of the researcher to the case under study makes him *an insider-researcher* (Adler and Adler, 1994; Unluer, 2012).

An insider researcher, in the broad sense, is a researcher who conducts research into his/her organisations or who belongs to the group under investigation (Breen, 2007). According to Rabe (2003)the position of the researcher as an insider or outsider in particular social research can be understood within three contexts: the context of power, the context of knowledge, and the role of the researcher in anthropology research. In this section, the context of knowledge is focused on.

One of the advantages the insider researcher has is background knowledge, which takes an outsider a long time to possess. Because the service being investigated is hosted in the same institution where the researcher is conducting his research and because of the involvement of the researcher in the process of reconstructing of services, the researcher has inside knowledge the outsider researcher may not possess. This includes contextual details about the services being studied, the internal configuration and setting of the former service system, the circumstances associated with failure of service, and the log resources of the ROAR system's internal processes.

The role of inside knowledge in this case study is in the demonstration and discussion of how open access analytics has been constructed, as well as the historical events that have occurred in the field of service operations. This role is supported by log analysis, case-effect analysis, and discussion. Additionally, the case study results are discussed in the light of open access repository

4.2.3.2 Log analysis

Log analysis is a powerful analysis technique that enables system designers and operators to investigate, detect, and troubleshoot system failures and evaluate the system performance (Jayathilake, 2012). Reconstructed OAI-PMH harvesting and data extraction tools are programmed to log any runtime exceptions executed during the harvesting and processing of the OAI-PMH metadata. These exceptions are indexed and stored in a MySQL database before being exported and analysed. This form of log contains vital information about the system interactions with the OAI-PMH end points on a large scale and provides insight into the types of technological issues that can constrain the typical use of the OAI-PMH framework.

4.3 The former ROAR analytics system

ROAR analytics is constructed on top of two components: the open access repositories index and OAI-PMH metadata aggregator provide harvesting support and act as a proxy server for other services utilising OAI-PMH metadata. Furthermore, the system is associated with four roles: repositories managers, registry managers, OAI-harvester managers, and ROAR analytics managers. The overall system is illustrated in Figure 4.1

RO1 Repository Manager: An extensive discussion about the role of a repository in the open access repository context is conducted in Section 5.3.3.2. However, concerning the ROAR system, their role is extended and associated with responsibilities such as indexing their repositories and managing the records in the ROAR registry.

RO2 Registry manager: An expert in the concept of open access who reviews the index records and maintains the data provided by repositories managers.

RO3 OAI-Harvester manager: A software developer who develop tools and maintains them to control the OAI-PMH harvesting process.

RO4 ROAR analytics manager: A data and analytics expert who designs and develops the ROAR analytics service.

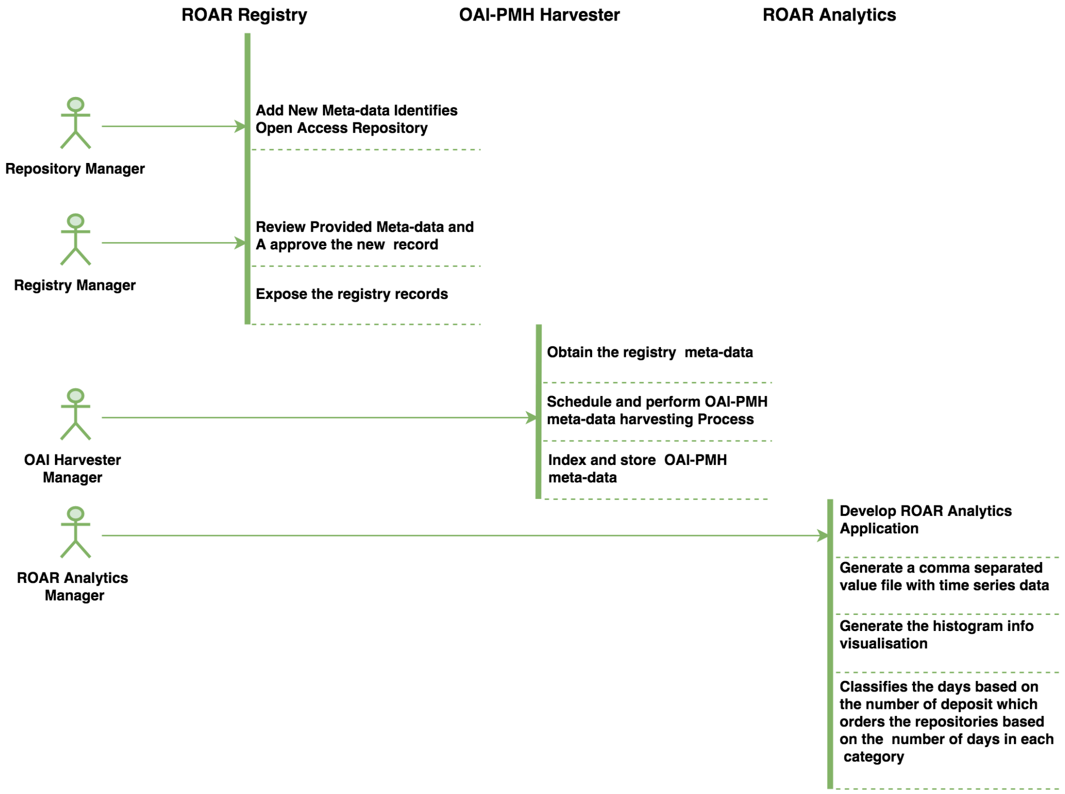

ROAR Registry
OAI-PMH Harvester
ROAR Analytics
Repository Manager
Add New Meta-data Identifies Open Access Repository
Registry Manager
Review Provided Meta-data and A approve the new record
Expose the registry records
Obtain the registry meta-data
OAI Harvester Manager
Schedule and perform OAI-PMH meta-data harvesting Process
Index and store OAI-PMH meta-data
ROAR Analytics Manager
Develop ROAR Analytics Application
Generate a comma separated value file with time series data
Generate the histogram info visualisation
Classifies the days based on the number of deposit which orders the repositories based on the number of days in each category

FIGURE 4.1: ROAR analytics former system interactions

4.3.1 The ROAR registry

In response to the requirements of the RQ1, the registry is maintained as a crowdsourcing platform that identifies open access repository nodes through the engagement of their managers. Thus, repository managers contribute general, geographical, technological, and OAI-PMH related details to the registry about their repositories. The crowdsourcing platform forms a network of repository managers who actively contribute to the identification process of open access repositories. Consequently, registry managers take responsibility for an expert-based filtering process over the repository managers' contributions. Thus, the review process prevents service misuse, enforces inclusion and exclusion criteria and evaluates the repositories' provided metadata.

4.3.2 The aggregator/OAI-PMH proxy

The registry underpins a centralised OAI-PMH metadata aggregator (OAI-PMH service provider), and a proxy provides the distributed OAI-PMH metadata from a single point of discovery. In the ROAR case study, the process is carried out by Celestial. Celestial aims to harvest the distributed OAI-PMH sources and provide the harvested collection from a single source. Thus, the tool mirrors the harvested metadata and provides it for other services, including ROAR analytics. Figure 4.2 demonstrates the ROAR system components, where the OAI-PMH aggregator is the core system that enables the extraction of the data required to generate a ROAR visualisation.

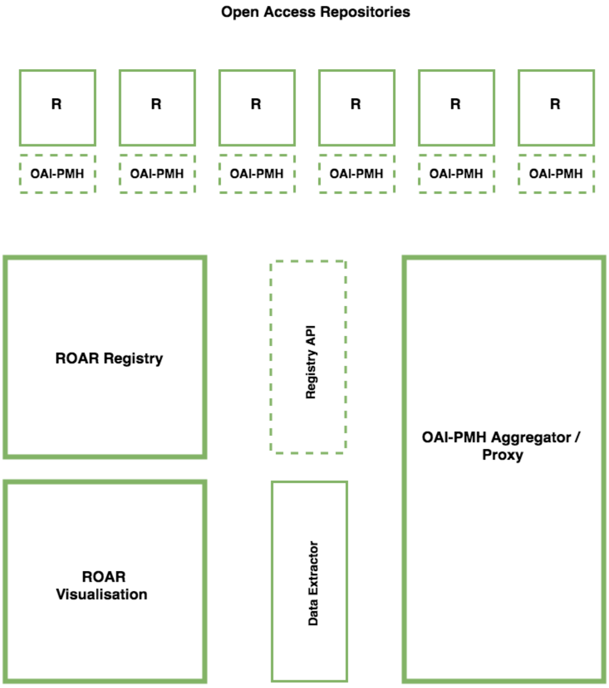


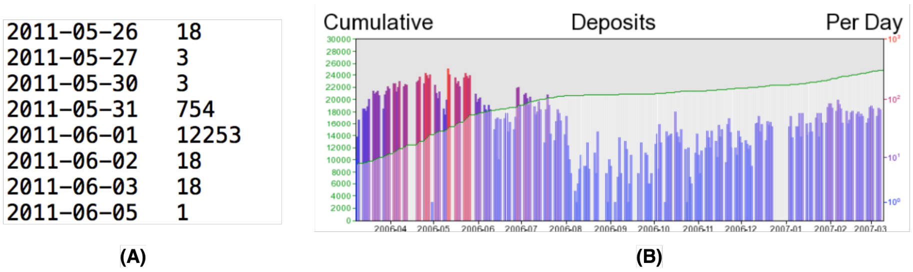
FIGURE 4.2: The ROAR former system is based on a central aggregator/proxy that harvests the OAI-PMH data providers' metadata and enables the extraction of the data required to generate a ROAR visualisation.

4.3.3 The ROAR analytics visualisation

ROAR analytics is a time series that represents a sequence of deposit activities in a particular repository within a specific period, and is presented as a histogram. The histogram presents the deposit activities with a daily interval granularity. Additionally, it is associated with a representation of cumulative deposited contents. ROAR analytics visualisations use in-house batch data processing tools to normalise the OAI-PMH metadata into a set of one-day time intervals associated with the number of deposit activities. The time intervals with no deposit activity are removed, as illustrated in Figure 4.3 (A). The figure shows a sample of the comma-separated value (CSV) data utilised to generate the ROAR histograms, which are illustrated in Figure 4.3 (B).

4.4 Open access analytics using OAI-PMH service provider: Analysis

This section goes through the process of ROAR analytics visualisation re-construction, which involves the examination of ROAR registry records, the operationalisation of



```
2011-05-26   18
2011-05-27   3
2011-05-30   3
2011-05-31   754
2011-06-01   12253
2011-06-02   18
2011-06-03   18
2011-06-05   1
```

Figure 4.3: Sample of (A) comma-separated value file and (B) ROAR histogram adopted from Carr and Brody (2007)

.

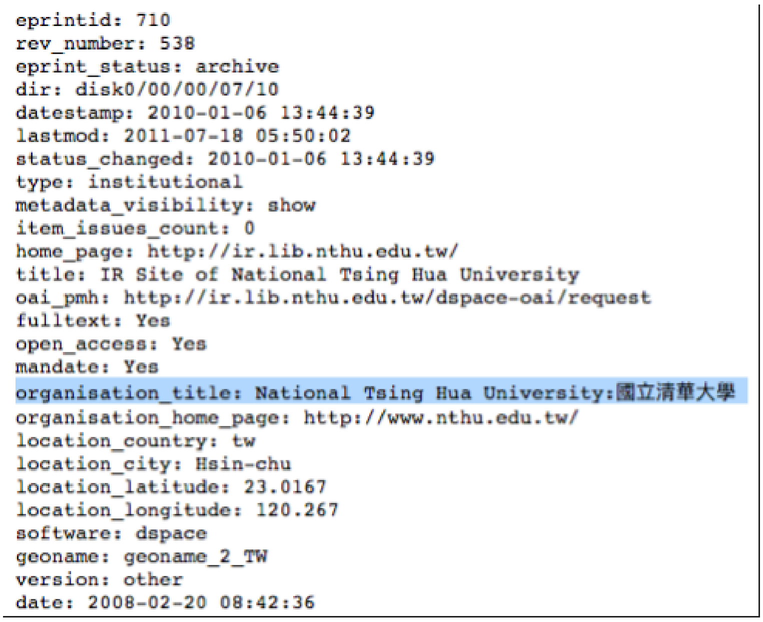
the 'deposit activity' concept, OAI-PMH metadata harvesting and processing, and the visualisation of deposit activities.

4.4.1 Discovery of OAR and coverage of ROAR analytics

The registry powers the analytics with a single point of discovery and enables a large scale of coverage. One of the objectives of ROAR analytics is to reflect in a timely way the growth of green open access on a worldwide scale. The large-scale analysis of open access repository deposit activities requires an approach identifying a distributed repository that complies with the open access publishing concept. As provided in the ROAR analytics system, the analytics use a registry indexes these repositories through a crowdsourcing platform to ensure a high level of coverage.

The ROAR registry is a collection of records comprising a set of attributes that catalogue open access repositories on a global scale. These attributes identify repositories based on their geographical, technical, historical, and performance details. The registry index is openly exposed through a variety of data representation protocols, such as MIME plain text and XML format. Additionally, the registry is integrated into other registries operated by the open access community, such as OpenDOAR, as well as integrated into other services such as ROARMAP. Figure 4.4 provides an example of the records represented in the MIME standard, identifying one of the Chinese repositories launched in February 2008.

The registry classifies repositories into eleven specific categories, in addition to a general category that accommodates unclassified repositories. Given this heterogeneity in the purpose and type of repository, the first challenge with analysing repository depositing activities arises because they cannot be analysed using a single data analytics approach given the variation of purposes, contents, and systems they adopt. The analysis of the varying classes of repositories may lead to an inaccurate interpretation of green open access adoption. For example, the research cross-institutional repository acts as an aggregator of open access literature on a geographical basis or subject basis, which may allow the replication of open access resources.



```
eprintid: 710
rev_number: 538
eprint_status: archive
dir: disk0/00/00/07/10
datestamp: 2010-01-06 13:44:39
lastmod: 2011-07-18 05:50:02
status_changed: 2010-01-06 13:44:39
type: institutional
metadata_visibility: show
item_issues_count: 0
home_page: http://ir.lib.nthu.edu.tw/
title: IR Site of National Tsing Hua University
oai_pmh: http://ir.lib.nthu.edu.tw/dspace-oai/request
fulltext: Yes
open_access: Yes
mandate: Yes
organisation_title: National Tsing Hua University:國立清華大學
organisation_home_page: http://www.nthu.edu.tw/
location_country: tw
location_city: Hsin-chu
location_latitude: 23.0167
location_longitude: 120.267
software: dspace
geoname: geoname_2_TW
version: other
date: 2008-02-20 08:42:36
```

Figure 4.4: Example of ROAR record exposed as MIME headers.

Furthermore, the ROAR index is not specific to open access repositories that are specifically designed to support the open access movement agenda and scope. It enables the indexing of a variety of types of repositories, including data repositories, web observatories, etheses, and e-learning resources repositories. However, the classification enables the registry user to filter and scale them based on their class. Given the fact that the non-research agenda repository is out of the scope of the open access repository concept, this thesis is more concerned with institutional repositories than the other types of open access repositories.

Although classification is clearly defined by the software system, the repository metadata are crowdsourced by the open access community, and the quality of the data is controlled using an expert review approach. However, the reviewing process fails to eliminate duplication because of the absence of a record duplication detection assistance tool. The OAI-PMH base URL is URI-compliant resource identification (Masinter et al., 2005), which should be unique to particular web resources. Hence, a duplication detection tool is used to identify duplication of records based on the OAI-PMH base URL. Even so, the duplication detection tools do not provide an optimum solution because some duplicated records exist with wilful purpose. For example, some repositories represent multiple institutions or departments; therefore, the repositories manager deliberately creates multiple records with a similar OAI base URL to represent different institutions. To understand and reveal the size of the duplication and the registry metadata quality issues, the registry entries are reviewed based on the following four criteria:

- The availability of the OAI-PMH base URL: Only entries with an OA-PMH base URL should be included.

- The uniqueness of the OAI-PMH base URL: Only entries with a unique OAI-PMH base URL should be included.
- The reachability of the OA-PMH base URL: Only entries with a working OAI-PMH base URL should be included.
- The association of the OAI-PMH base URL with XML resources: Only entries with an OAI-PMH base URL for XML metadata should be included.

The quality of open access registries can be improved by the integration of other open access registries for them to take advantage of each other. The availability of OpenDOAR ID is harnessed to integrate the ROAR registry entries and OpenDOAR entries, with the assumption that the OpenDOAR is more carefully maintained than the ROAR registry.

An examination of the availability of the OAI-PMH base URL scales down the number of registry entries from 4,326 records collected from the ROAR registry on 1 October 2016 to 3,360 entries with the OAI-PMH base URL. Because the registry records are not continuously reviewed and target distributed infrastructure with continuous changes to the node, the OAI-PMH has to be validated prior to any harvesting process. The initial validation shows that only 1,617 entries have a valid link. Furthermore, the majority of the invalid OAI-PMH interfaces point to unreachable destinations or direct users to an HTML page.

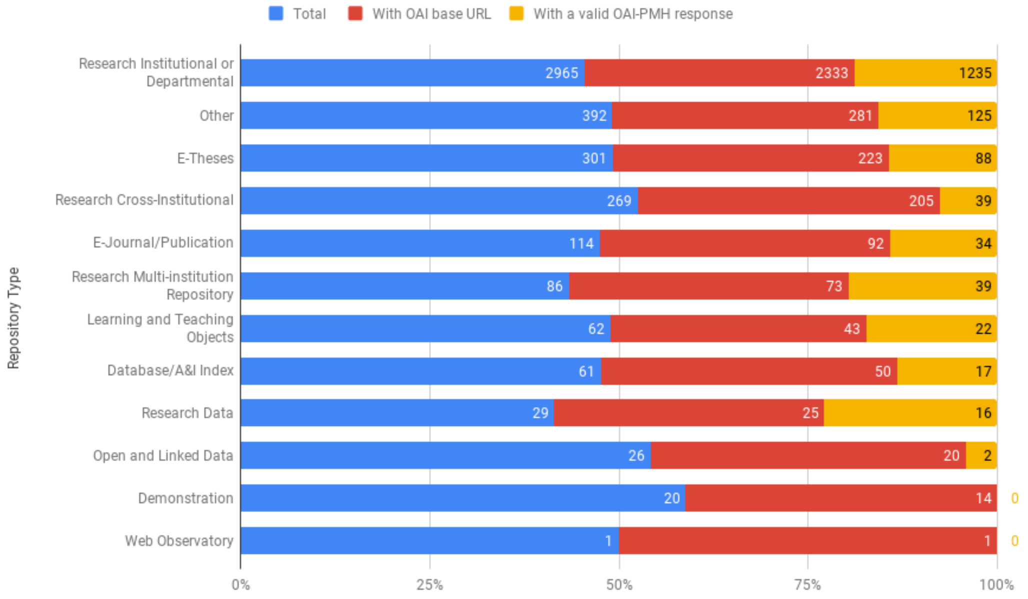


FIGURE 4.5: The status of the registry entries associated with the types of repositories.

Although the use of an OAI-PMH base URL to check entry duplication can prevent the replication of an open access repository and reduce efforts to harvest and process repository metadata, it may make the links between registry entries and visualisation more complex in later phases. Thus, the duplicated entries are kept and harvested, and their metadata are processed and visualised. Figure 4.5 shows the distribution of ROAR entries to repository types before and after the examination of registry entries.

4.4.2 The operationalisation of 'deposit activity'

Based on the second objective of ROAR analytics, it is necessary for the analytics to quantify the deposit activities within an individual repository. The open access repository uses the OAIS functional reference model as a general framework to conceptualise the open access repository functions at the conceptual level. One of these functions is the ingest function. According to CCSDS (2012), the ingest function is the acceptance of a submission of information packages from the producer and the preparation of an archival information package and its descriptive information, which can be used later for the generation and dissemination of the information package to the consumer. Deposit activity can be defined as the execution of the ingest function. However, in terms of open access publishing, deposit activity is the availability and public accessibility of full-text research on the Internet (see Section 2.1.1.3). Therefore, deposit activity can be operationalised by the dissemination of full-text research using open access repositories and through the use of its dissemination methods, including the OAI-PMH framework.

Within the OAI-PMH, the OAI Identifier uniquely identifies each item. The OAI identifier should comply with the URI syntax (Masinter et al., 2005). The individual records are encoded in XML and divided into two main sections: the header section comprises the OAI Identifier, the sets the item is related to, and the datestamp; and the metadata section, which is where the item metadata is exposed as a variety of metadata schema such as Dublin Core and MARC. Each resource hosted in the repository is uniquely identified by the OAI Identifier. Thus, the number of OAI Identifiers exposed by the OAI-PMH can reflect the number of items within a particular repository. Further, the datestamp associated with each OAI Identifier can reflect the date and time of the deposit activity. Although the header blocks are standardised and provide an easy option to calculate deposits in a particular repository and for a specific time using an OAI Identifier and datestamp, there are substantial limitations associated with using the datestamp as an indicator of deposit data and time.

The value of a datestamp is not in the data and time of an item's creation, but instead in the last modification date and time. Thus, its value resets after any modification to the item entries in the repository, which may lead to an inaccurate reflection of the date and time of deposit activity. This issue can influence the deposit activities that are modified after they have been created. However, it can also influence the whole collection because of a set of managerial actions that take place as a part of the repository management. For instance, the repository managers recover their data and all of the records exposed with a new datestamp pointing to the date and time of recovery.

One of the examples of this issue can be seen in the case of the Southampton Eprints open access repository (see Figure 4.6). As the chart shows, an initial deposit was made in March 2014, but the construction of the Southampton E-Print repository took place at the end of July 2003. The former implementation of ROAR analytics operationalised the deposit activity date and time by the dedication of the first date and time associated with a particular OAI Identifier, which implies the need for a cumulative harvesting and examination of the OAI Identifier as batch processes taking place before any visualisation processes.

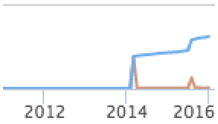


FIGURE 4.6: The managerial activities significantly influenced the item date stamp values (Example: Although the University of Southampton Institutional Repository was established in the second quarter of 2003, the figure shows the initial deposit activity was in the first quarter of 2014).

The analysis of open access deposit activity cannot be fully operationalised without the detection of full-text deposits instead of the availability of metadata. The detection of full-text resources using the OAI-PMH is a complex process that requires large-scale analysis because of the variations in the adoption of OAI-PMH and because of how the repository managers use and interpret the metadata schemas in their implementation. Additionally, the resources are out of the scope of OAI-PMh implementation (Van de Sompel et al., 2004). However, the operationalisation of repository deposit activities using OAI Identifiers and analysis of the deposit dynamic can shed light on the dynamics of green open access (Carr and Brody, 2007).

4.4.3 The harvesting and processing of OAI-PMH metadata

The OAI-PMH is a low barrier interoperability protocol based on metadata harvesting. The framework enables interoperability between two main entities (Lagoze and Van de Sompel, 2003):

I. The data provider where the OAI-PMH protocol is implemented and used to expose metadata about their managed resources

II. The service provider that harvests the exposed metadata to construct added value services on top of the OAI-PMH metadata

The OAI-PMH empowers data providers with distributed resource discovery using metadata (Van de Sompel et al., 2004). This metadata is provided as a set of records that have a set of attributes with a specific metadata schema in an XML-encoded byte stream associated with an item identifier that identifies resources hosted in the repository. The OAI-PMH data model is illustrated in Figure (4.7). It represents the hierarchical relationship between the resources, items, and records and with the metadata schema format.

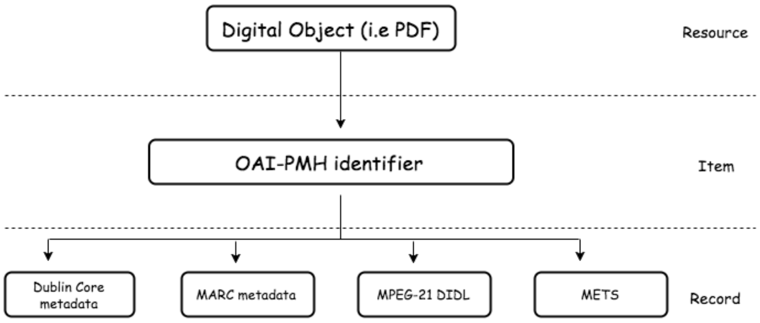


FIGURE 4.7: OAI-PMH data model representing three tiers of OAI-PMH data redrawn from Van de Sompel et al. (2004)

.

The OAI-PMH data flow between the harvester and data provider which is (the harvester) a client application issues the OAI-PMH request to harvest the OAI-PMH metadata from the data provider's end. Within the data provider, the request is processed in the repository by a server designed to handle OAI-PMH requests and respond to harvesters with the requested metadata. The OAI-PMH uses a resumptionToken, a string value used to continue harvesting incomplete lists in a sequential manner, and a fault tolerance approach to deal with errors and recover the harvesting process. Furthermore, it supports selective harvesting using a datestamp that represents the last changed date and time. Hence, selective harvesting allows the service provider to harvest the item's metadata that have been added or changed between the two timestamps given in the OAI-PMH requests.

The OAI-PMH data provider nodes expose their metadata as chunks of 100 records or more than 200 records supported with resumptionTokens employed to navigate through record chunks. These resumptionTokens allow the harvester to recover from a network failure. However, only the next resumptionToken is available because instead of providing all resumptionTokens in the first request constraint, the harvester follows a sequential harvesting approach within each individual repository. This implies that the harvesting process requires a massive number of HTTP requests, making it a time-consuming process.

A large number of repositories intended to be harvested necessitate the utilisation of available solutions that may accelerate harvesting activities. In this case study, one machine with one network interface is used, which causes a network bottleneck problem. The amount of time spent by the harvester waiting for the data provider's reply can be exploited by the harvester to deal with another request within the same repository or another repository. This technique was proposed by Suleman (2006) to parallelise the harvesting process within a single repository by operating multiple harvester nodes to synchronise the process and share the network resources. However, the adoption of this form of harvesting is constrained by procedures adopted by repositories to limit the number of OAI-PMH requests within a time frame. Similar issues were reported by Brody (2006), even though, in the case of harvesting multiple repositories, this limitation does not apply. The process is parallelised to harvest multiple repositories at the same time. In this case study, the OAI harvester is a Java tool developed to perform OAI resources retrieval and pre-processing. Therefore, the Java concurrency framework is used to implement the parallelisation of the harvesting process (Goetz, 2006).

The harvester retrieves OAI metadata, which are then serialised into XML files that can be stored and maintained by native XML databases (NXD) (Jagadish et al., 2002), such as the BaseX[1] database management system, where XML online analytical processing (OLAP) system can be adopted to process XML files on demand (Park et al., 2005). However, in the case of distributed infrastructure that is implemented and maintained by different administration domains, the quality of the adoption of OAI-PMH specifications vary, requiring a validation of XML data. Otherwise, the NXD functionality is interpreted.

On the other hand, XML files can be parsed during the harvesting process and ingested into a relational database management system to take advantage of the mature Relational Database Management System (RDBMS) . XML parsing can take place using the built-in programming language or using third-party frameworks such as Document Object Model (DOM), Simple API for XML(SAX), Java Document Object Model (JDOM), Streaming API for XML (StAX), XML Path Language (XPath), and Flexible XML framework for Java (DOM4J). Haw and Rao (2007) benchmarked four of the common XML parsing frameworks, finding the SAX-based parser (Xerces) to be better in performance than the DOM-based parser (xParse). However, it is not just performance that matters during the parsing process, the scale and type of requirement may be better parsers for a particular case. The drawback of parsing metadata during the harvesting process is the loss of records during the process. The harvester may encounter a number of syntax errors as a result of the low adoption of OAI-PMH specifications. This initial harvester may use this approach to harvest and process OAI-PMH metadata, enabling the harvester to collect exceptions encountered during the harvesting and processing of XML metadata. These exceptions are logged and analysed to acquire insights into the scale of these exceptions and the types of errors encountered during the process (see Figure 4.8).

Another approach to dealing with the harvested OAI metadata is the XML-enabled database, which is a relational database management system (Pardede et al., 2008) but is supported by XML parsing functionality. This approach reduces the need to

[1] http://basex.org/

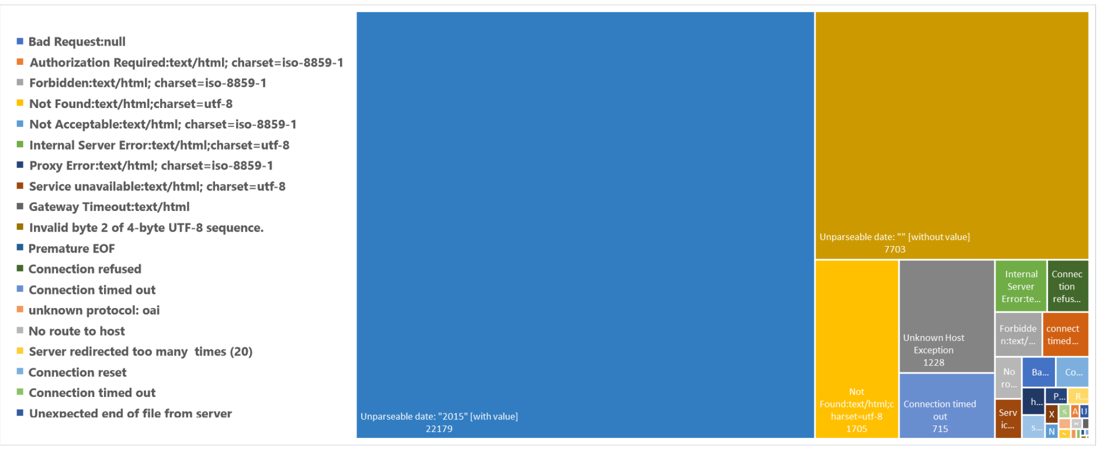


Figure 4.8: Exceptions encountered during OAI metadata harvesting and processing.
.

parse the OAI-PMH metadata during harvesting as well as takes advantage of RDBMS functionalities to maintain and retrieve the OAI-PMH metadata. However, the process of query and retrieval of the OAI-PMH metadata becomes complicated because each file is composed of a number of OAI items (ex:100) and is stored in a single entity in the database. In light of this problem, this approach can be incorporated into the XML segmentation technique by identifying regions and records within a particular XML file (Zhai and Liu, 2006). This technique enables the storage and maintenance of OAI-PMH by their original OAI Identifier without parsing the whole file; instead, each item record is stored in the RDBMS in its XML format (see Figure 4.9). The second version of the harvester is moved to this approach.

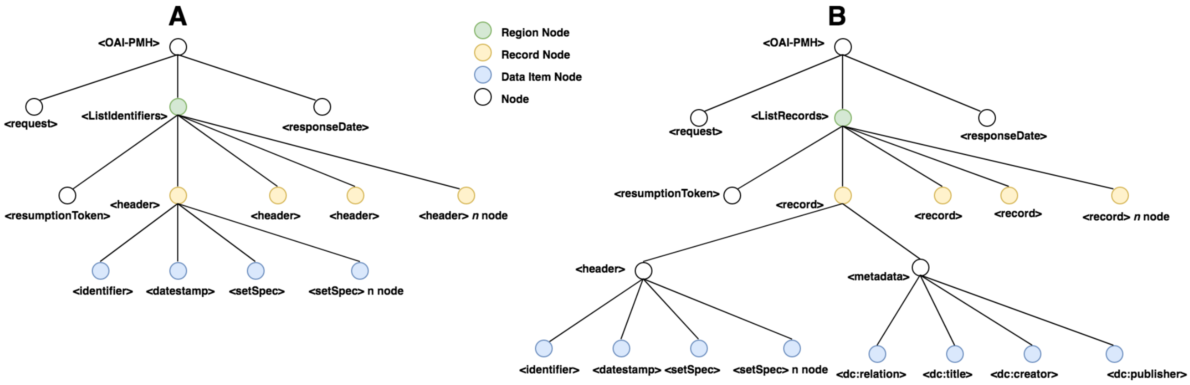


Figure 4.9: OAI-PMH meta-data segmentation and its move from a tree-based structure to a record-based one: A) ListIdentifiers response and B) ListRecords response.
.

Although there are many powerful solutions for pre-processing of unstructured and semi-structured data, the MapReduce framework offers a functionality that can help overcome the challenges associated with OAI-PMH processing. The RDBMS is used with the assumption that the level of standardisation in OAI allows the transformation of resources from a semi-structured to a structured format and reduces the pre-processing activities performed in the MapReduce framework (García et al., 2016). It has already been adopted by projects targeting the same resource (Wu et al., 2015).

Last, the scale of open access analytics requires harvesting a large number of records and acquiring a large volume of data; this in turn requires more processing resources and capabilities. In this case study, more than (44) million items were harvested and processed to enable lightweight analytics visualisation, empowering the community with a simple but significant insight into the development of open access repository adoption and green open access publishing.

4.4.4 Visualisation of Open Access Repository Deposit Activities

The ROAR visualisation is a time series analysis. The time series is a set of sequences of time point provided with a given base granularity, such as seconds, minutes, days, months, and years. These time points are given in continuous sets of 'time intervals'. The raw data provided by OAI metadata represent a set of events extracted from item identifier attributes and their deposit date and time extracted from the item timestamp (the first timestamp associated with the item identifier).

Thus, in a single repository, there is a set of deposit activities. Each deposit activity is associated with a timestamp. To map each deposit activity to the time series, the relevant time point should be identified using a timestamp. However, they have to be provided with similar base granularities. The OAI-PMH datestamp is provided in UTC (Coordinated Universal Time) datetime format (Wolf and Wicksteed, 1997). For instance, in (2017-07-17 T 15:46:09 Z), the first part represents the date $(2017-07-17)$ in the $(YEAR - MONTH - DAY)$ format, and the second part represents the time $(15:46:09)$ in the $(HOURS : MINUTES : SECONDS)$ format. Therefore, several base granularities can be extracted from it. For example, a day base granularity can be represented as follows: $\langle 17, 07, 2017\rangle$,$\langle 15, 07, 2017\rangle$, $\langle 10, 06, 2017\rangle$.

Hence, for each item to be assigned to a particular day base granularity time interval, the time point of its datestamp should be a time point in the time interval. In ROAR analytics, a day base granularity time interval is used, and the relevant activities are calculated. This process is calculated using a normalisation process that assigns a numeric value to each time interval based on the relevance of the item to the time interval.

Therefore, ROAR visualisations are generated using normalised data on each repository deposit activity provided as flat files to be processed by the visualisation app. In this case study, the normalisation involves scaling and data reduction techniques that summarise the repositories deposits activities (Patro and Sahu, 2015). Hence, the normalised data should provide the visualisation app with a monthly deposit count associated with the repository ID in ROAR. Normalising the OAI-PMH metadata using real-time processing is a resource-demanding process given the massive number of records that need to be processed in each normalisation task. Furthermore, the dynamic of OAR in roar analytics is calculated in a daily and monthly manner, which does not require real-time processing. Consequently, batch data processing is used to normalise deposit activities.

Technically, visualisations use statistical reports stored as JSON files that are ready to be read by the visualisation app. The visualisation app uses Highchart API[2], an interactive javascript framework that gives developers a set of visualisation functions. The Highchart API is used to develop web apps using PHP[3] scripting language to read JSON files and generate interactive visualisation. Figure (4.10) gives an example of charts generated by the visualisation app on the University of Manchester's deposit activities.

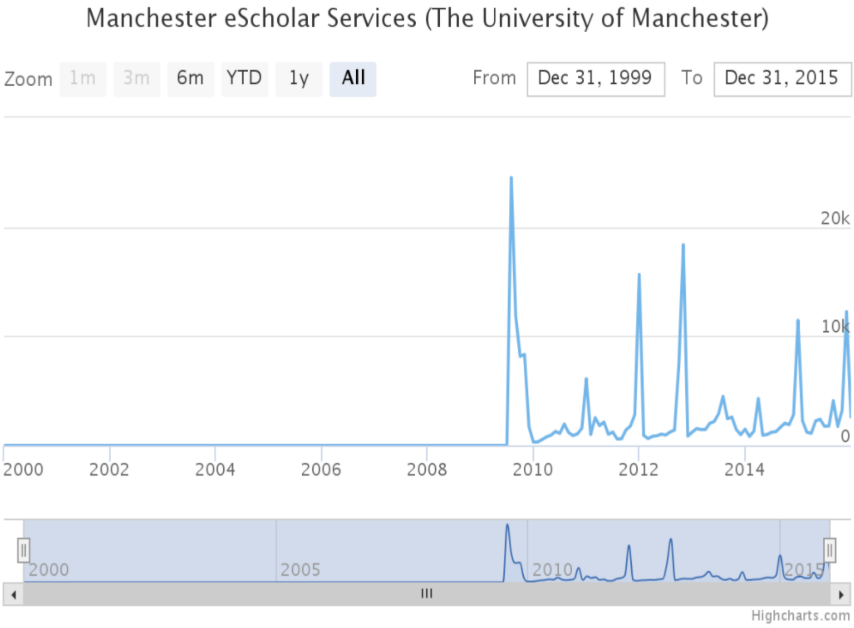


Figure 4.10: Example of charts generated by the visualisation app on the University of Manchester's deposit activities.

.

4.5 Discussion

The open access repository infrastructure is designed as a distributed infrastructure, which prevents it from the single point of failure issues. The services providers that aggregate the repositories metadata are developed as a centralised system collecting, processing, and indexing repositories metadata to enable value-added services. This

[2] https://www.highcharts.com/docs/index

[3] https://www.php.net/

makes them vulnerable to single points of failure. Although, the exchange of metadata is the mean instead being the purpose that OAI-PH is designed to support . The OAI-PMH is designed to support resources hosted in the distributed repositories. However, in the case of ROAR analytics, the metadata are the data that are used to carry out the analysis. Therefore, the failure of the ROAR system and the loss of the metadata can suspend the service.

System failures can lead to unrecoverable data loss. The system loses a massive amount of data collected cumulatively through the continues harvesting of repositories metadata on a global scale. Because of the aforementioned issues associated with datestamps, ROAR visualisations are misrepresented and mislead if the analytics is reconstructed with the same strategy and with a new harvesting process. One of the solutions to this dilemma is to use other service provider collections. Therefore, the CORE (COnnectingREpositories) API is examined to determine the functionality of its collection for ROAR requirements.

The CORE system provides three levels of access to its data (Knoth and Zdrahal, 2012): 'access at the granularity of papers', 'analytical access at the granularity of collections' and 'access to raw data'. Access to raw data is used because it provides unprocessed metadata and can work as a proxy that maintains different data points in the repository lifetime. Nevertheless, the CORE data are not harvested in the same interval within a single repository or cross-repositories. This issue can influence the consistency of analysis within a single repository or cross-repositories.

This type of failure draws attention to service provider sustainability, especially were designed for analytical purposes. The open access repositories research community pays special attention to institutional repositories' sustainability (Nkiko et al., 2014; Bradley, 2006; Chan et al., 2005) as its core agenda for preservation research. On the other hand, service provider sustainability has been given less attention. It is discussed and approached by forming global, regional, and nationwide communities beyond funding projects. Lossau and Peters (2008), in their work on the Digital Repository Infrastructure Vision for European Research (DRIVER), stated,

> The DRIVER community is sustainable in the form of a Confederation, in which the members assume responsibility for the DRIVER objectives, contribute data, offer services, share expertise and suggest strategic direction to the DRIVER community. These objectives form strong natural incentives, beyond project funding, to build an international repository community that is firmly embedded in national communities and supported by significant alliances with LIBER, SPARC Europe and eIFL. The DRIVER Confederation is envisaged as the permanent organisational backbone of the DRIVER Community. (Lossau and Peters, 2008, p 445)

Thus, service providers are implemented as a distributed service architecture provided as a nationwide infrastructure or regional-wide infrastructure. For instance, Müller et al. (2009) describes the Open Access Network project, which aims to support open access infrastructure for Germany. This infrastructure adopts architecture that supports value-added services and empowers developers with a testing development environment that enables easy involvement in the development of value-added services, including analytics services. Accordingly, the long-term management and establishment cost is paid by nationally funded infrastructure.

In 2016, the Confederation of Open Access Repositories (COAR) launched the next-generation open access repository working group with the following vision:

> To position repositories as the foundation for a distributed, globally networked infrastructure for scholarly communication, on top of which layers of value-added services will be deployed, thereby transforming the system, making it more research-centric, open to and supportive of innovation, while also collectively managed by the scholarly community.

The vision indicates a strong emphasis on supporting a new form of value-added services supporting the scholarly community in general and open science in particular. Therefore, at the end of 2017, in their report 'Next Generation Repositories Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group', COAR made recommendations for the adoption of a number of technologies and standards that can expand the types and numbers of value-added services (see Figure 4.11) that can be operated and supported by open access repositories infrastructure, including analytics practices (Rodrigues et al., 2017). However, these recommendation and technologies will require time to be adopted on a large scale.

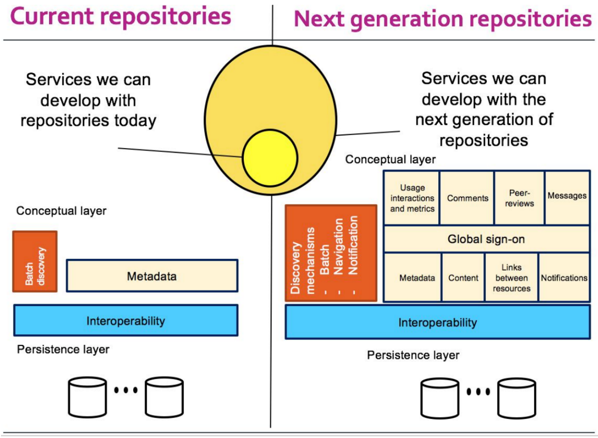


FIGURE 4.11: The vision for next-generation repositories and the emphasis on value-added services; adopted from Rodrigues et al. (2017)

.

4.6 Chapter Summary

In this chapter, the delivery of open access analytics visualisation using the OAI-PMH service provider is demonstrated through an explanatory case study. ROAR analytics takes advantage of open access registry to identify, discover, and de-fragment distributed open access repositories nodes, which are subject to entry quality issues. Furthermore, the service incorporates a centralised OAI-PMH service provider to harvest and process the OAI metadata on a monthly basis to generate time-series visualisations representing deposit activities. Through this process, a set of challenges are discussed, including low OAI-PMH adoption, processing and network bottlenecks, big data volumes, long-term management, and service provider sustainability. These can act as a barrier to harnessing repositories data for open access analytics practices.

Chapter 5

The OAA-OARD-SM Conceptual Framework Development

5.1 Introduction

The first research question aims to provide an understanding of to what extent the conventional data-service provider approach, along with the OAI-PMH interoperability protocol, can support the requirement of open access analytics. Thus, ROAR analytics is used as an explanatory case study to demonstrate the open access analytics process. In addition to the demonstration of the process of open access analytics as an added value service on top of OAI-PMH infrastructure, the case study demonstrates that a set of challenges and limitations increase the complexity, as well as the cost of the exploitation of the open access repository data value for open access analytics practices.

These set of challenges and limitations position the case study of ROAR analytics as case provides motivation for an investigation on a new form of collaboration and realisation of the data analytics process that may pave and reduce the effect of these challenges. The second research question was derived to evaluate the opportunities offered by the social machine concept with regard to the realisation of a particular system process (the open access analytics process) using a web-based socio-technical system. Thus, the following research questions was created:

> RQ (2): Is the concept of social machines useful to re-conceptualise the process of open access analytics within an open access repositories ecology?

The aim of this research question is to acquire an understanding of open access analytics using open access repository data as a social machine process from the conceptual level of abstraction. Consequently, three main objectives were determined to achieve this understanding; I) the systematic acquisition of the relevant literature across the three main research areas, specifically the open access repository, data analytics and the social machine literature, II) the conceptual examination and the analysis of the process of open access analytics and III) the conceptual synthesis of the process of open access analytics as a social machine process within a single conceptual framework.

5.2 Web Science and the Lens of Social Machines

Web-related aspects are studied and investigated by several research communities, including computer science, economics, social science and human–computer interaction. However, their pragmatic approach limits the web to a delivery vehicle for content or dynamic content that enables social interaction, instead of the primary object of the study (Hendler et al., 2008). To overcome this limitation, web science is created to study the web as a whole, a phenomenon and engineering artefact. To achieve this vision, the web science research community draws attention to contributions attempting to capture web properties using (for instance) network theory and actor network theory to spectate their limitations in capturing the web properties and align them with the web science core agenda: studying the web as ecology (ibid). However, these approaches present limitations to capture and interpret the blend of web properties, such as combinations between the technical and web aspects of the web.

Accordingly, the web science research community drew attention to the concept of social machines to understand the web as evolving artefact. The social machine concepts provide researchers with a lens to view the web-based socio-technical system as a process co-created and released by humans and machines as participants (Smart et al., 2014). Hence, both the technological and human elements are treated equally with respect to the realisation of the system process (ibid). This concept was used to interpret and understand the web itself and the system within the web (Smart et al., 2014; Hall and Tiropanis, 2012; Luczak-Roesch et al., 2016). Therefore, the correlation between web science and social machines is deep in their agendas and core principles. Hall and Tiropanis (2012, p 3863) define web science as "the theory and practice of social machines".

Based on the perspective of web science pioneers and its creators, web science is not just synthetic science concerned with modelling the current web. It also involves the synthetics in engineering a new infrastructure and the creation of new beneficial systems, as well as powerful tools and analytics in understanding how humans use and contribute to it (Berners-Lee et al., 2006). Therefore, the social machine lens is used to understand and reconstruct various forms of real-life practices that exist in the web, including (but not limited to) smart cities (Ahlers et al., 2016), governance (Tiropanis et al., 2014b), and e-learning (Arafat et al., 2019).

One of the phenomena being interpreted as a social machinery process is open access publishing (Smart et al., 2014). In open access publishing, the open access repositories are the technical infrastructure supporting the polling of open access e-prints worldwide, thereby achieving an overall agenda of open access movement and the immediate removal of access and legal barriers placed by commercial publishers. In addition to the e-prints level, there is an established repository level process based on the crowdsourcing approach to de-fragment the open access repositories to provide a single virtual repository. This enables the discovery of the repositories and, accordingly, their resources from a single access point.

As discussed earlier in Section 2.2.4 and demonstrated in Chapter 4, open access analytics takes a similar approach and is constructed as an added value service on top of open access repositories aggregators, in which the open access repositories play a passive role in the process. Accordingly, it is speculated here that the actors in the open access analytics process can be reframed through the lens of the social machine. Hence, the entities and actors surrounding the process can play the role of participants instead of users. The concepts of social machines, data analytics, web observatory and open access repository provide not only a powerful theoretical framework to understand open access analytics in novel approaches but also the context, methods and tools to power the open access communities with new aids.

5.3 Development of the OAA-OARD-SM Framework

The qualitative research and conceptual analysis requires a high level of transparency during the process of reporting its procedures, as well as its findings and interpretation (Creswell and Miller, 2000). Thus, this section discusses the development process of the conceptual framework. Thus, it highlights the flow of development and provides justifications for the development producers. On the other hand, the output of development process is provided in Section 5.4.

5.3.1 Phase (1): Retrieval of the Related Literature

This phase aims to systematically acquire a comprehensive source of knowledge regarding the process of open access analytics and the concept of social machines though a set of procedures, utilising scholarly databases services.

5.3.1.1 Step 1: Identifying the Research Areas

The conceptual analysis examines the open access analytics delivery process using pen access repository data with a direct focus on the lens of social machines. Therefore, three research areas have been identified as a source of knowledge, namely open access repository, social machine and data analytics.

5.3.1.2 Step 2: Define Research Areas Roles

The role of each research area corresponds to open access analytics and the goal of the conceptual analysis is determined. According to Maxwell (2013), this process is important to avoid a 'covering the field' strategy, instead of focusing on what is relevant to the topic being studied. The research areas selected represent three main roles: context, model and process. The open access repository is identified as the context, and the social machine as a model that needs to be adapted to support the data analytics process.

- **Context**: the term context is defined by Rodden et al. (1998) as the environment or the situation around the issue being investigated. Therefore, the context of a task is the set of circumstances around it that are potentially relevant to completion of the task (Hull et al., 1997). This definition of context is adopted from context-aware application research community, which can be used as a general perspective of what the context is. Given the context of the analytics process, the concepts around the open access repository, processing within the repository system and the concepts influencing the uses of open access repository data should be part of the conceptual framework. The framework is designed to cope with a variety of analytic applications within the umbrella of the open access analytics concept. Therefore, the framework is not specific to a particular agenda, however, it is specific in terms of the data source (the open access repository data) and the analytics subject being analysed (open access publishing).

- **Process**: the term process refers to a set of actions or steps taken to achieve a particular end. The design research community uses it to denote the abstract nature and conceptualisation of these actions. Eckert and Stacey (2010) defined it as:

  > an abstract and relatively general conceptualisation of what is done or should be done to design a product, that is meant to apply to a class of products, an organisation, or an industry sector. (Eckert and Stacey, 2010, p 4)

  While the open access repository research community have made a number of efforts in analytic applications (Knoth and Zdrahal, 2012), these efforts have focused on the infrastructure and system design to deliver the analytic services, as well as open access repository data, using the standards adopted by the community for dissemination purposes such as OAI-PMH. However, the social machine model draws attention to the collaborative task as a process, instead of its technological infrastructure. Thus, the role of data analytics research is to inform the study on analytics as a process from conceptual level instead of infrastructure or systems.

- **Conceptual model**: the conceptual model is a subset of a theory, which consists of a set of concepts and relations and represents the abstraction of a phenomenon or physical object (Routio, 2007). The social machine is an emerging concept and its attributes and features provide a model that can be used to design or understand the interactions of social machines. Therefore, the role of the literature is to create a lens to synthesise the process of open access analytics as a social machinery process.

5.3.1.3 Step 3: Identifying the Links between the Research Areas

Although these research areas are divided into their research focus and directions, there are research topics that overlap. These topics were identified in order to determine the position of the framework in relation to the areas determined in Section 5.3.1.1, and to uncover the overlapping sub-areas. It draws special attention to I) 'Web Observatory' concept as a social machine model that supports the data analytics, II) the added value services as an existing utilisation of analytics within the open access repository community and III) open access repositories as social machine that collects and preserves open access resources. The overlap between the research areas, their role and the position of OAA-OARDA-SM conceptual framework to these research areas are illustrated in Figure 5.1.

5.3.1.4 Step 4: Define Search Strategy and Eligibility Criteria

The literature related to the three research areas was gathered using the Scopus scholarly database to achieve a good coverage and ensure the literature was searched to a high standard. Also, the literature search process was carried out taking into consideration the role of each research area in relation to the framework development process, as highlighted in Section 5.3.1.2. Scopus as a database is more advantageous compared to other well-known databases such as Web Of Knowledge (Falagas et al., 2008; Mongeon and Paul-Hus, 2016) in terms of literature coverage, as it contains more than 21,000 journals covered by 5^{th} of October in 2018 (Pellack, 2018).

With regards to an open access repository, in addition to open access repository being a keyword, three other keywords are used interchangeably in the literature to denote this, specifically: 'institutional repository', 'trusted repository' and 'digital repository'. These keywords are used to identify an open access repository in the relevant literature. Although these keywords are not specific to analytic activities using open access repositories, the term was used due to the need to achieve high recall. Based on the role of open

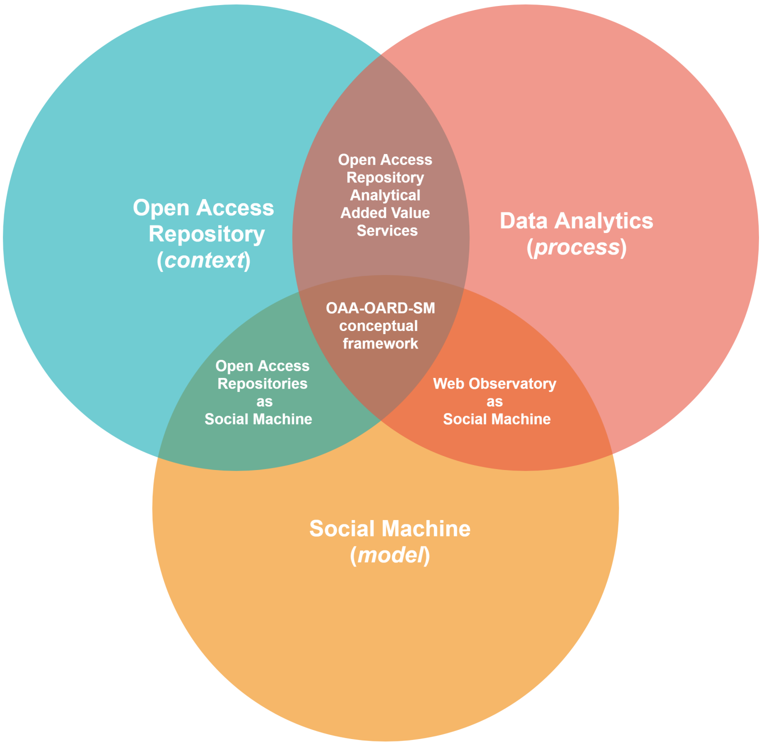

Open Access Repository (context)
Data Analytics (process)
Social Machine (model)
Open Access Repository Analytical Added Value Services
OAA-OARD-SM conceptual framework
Open Access Repositories as Social Machine
Web Observatory as Social Machine

Figure 5.1: Position of OAA-OARD-SM to the areas contributing to its development.

access repository literature in this process, a high recall is significant. The role of this literature is to uncover any context related concepts that may influence the process of open access analytics using open access repository data. Nonetheless, the use of generic words usually generates a large collection that requires manual processing. However, it is feasible in a situation where the retrieved records are within an applicable range to process them manually. The retrieved records are filtered using a set of exclusion and inclusion criteria.

In this research, two forms of eligibility criteria were used: general criteria were applied to the three research areas and specific criteria were introduced for each of the three targeted areas. While the generic criteria included two predetermined ones (Table 5.1), the specific criteria (Tables 5.2 and 5.3) emerged during the manual processing of the retrieved literature using the study titles and abstract metadata provided by Scopus.

The data analytics literature was approached by the extension of the data analytics concept by identifying the themes related to data analytics (Cooper, 2012a) including knowledge discovery and data mining, information visualisation and visual analytics (see Section 2.1.2.3). Hence, the keywords were extended to cover these activities to include the following: ‘information visualisation’, ‘knowledge discovery’, ‘data mining’ and ‘visual analytics’ (C5). Based on the role of data analytics within the conceptual framework development, the ‘process’ phrase is concatenated to these terms to recognise the literature regarding the conceptualisation of the analytics process (C6).

C-NO	Criterion	Justification
C1	The work should be available in English or Arabic.	Limitation of the researcher to process other language resources.
C2	The work should contain or discuss the concepts of open access repository data, analytics or social machines or present attributes or items of related concepts.	To align with the definition and attributes of the conceptual framework discussed in Section 5.3.1.1.

Table 5.1: The general criteria utilised to filter the contributing research areas literature

C-NO	Criterion	Justification
C3	The work should discuss, review or introduce the relevance of a concept or influence the open access repository as a data source for the analytics process.	This is based on the goal of the framework. In addition, there are a number of works that refer to *"open access repository"* not as the topic of the research, but instead as a source of literature to conduct the work on.
C4	The work should be aligned with the scope of the open access movement discussed in Section 2.1.3	To limit the scope to the open access movement target resources. However, the open access repository content is not limited to open access research, but instead a heterogeneous resource. Hence, the works that mainly focus on educational resources and research data curation were excluded.

Table 5.2: A specific criteria utilised to filter open access repository literature.

C-NO	Criterion	Justification
C5	The works with a direct focus on the three themes of data analytics, namely: data mining and knowledge discovery, information visualisation and visual analytics are included.	A considerable efforts placed by these research communities to understand the analytics process, where a set of process conceptual models are proposed to illustrate the analytical process from vary perspectives and themes.
C6	The work should be associated with the conceptualisation of the process of data analytics.	The majority of the work presents data analytics as a system or architecture which is not aligned with the aim of this study, as the social machine is process centric. Also, the aim of the framework is to highlight the process, instead of the systems or architecture.

Table 5.3: The specific criteria utilised to filter data analytics literature.

In contrast to the former two research areas, the literature related to the social machines is carried out with a set of pre-selected seminal works that consolidate the community efforts related to the concept of social machines in a single theoretical framework. These efforts eliminate the ontological ambiguity of the concept of the social machine, as well as enables the adoption to be aligned with one of the perspectives of social machines (the various perspectives of what is social machine means are reviewed in Section 2.3.1). Using a similar approach, the web observatory literature is examined.

During the examination of the open access repository literature, four types of resources were distinguished: I) a referential model, II) a conceptual framework or model, III) empirical studies and IV) value-added services and systems. The reference model is a road map that operates from a higher level of abstraction to provide an understanding of the abstracted environment and its relationship with other entities. Indeed, the CCSDS (2012) defines the reference model as:

> a framework for understanding significant relationships among the entities of some environment, and for the development of consistent standards or specifications sup- porting that environment. A reference model is based on a small number of unifying concepts and may be used as a basis for education and explaining standards to a nonspecialist. (CCSDS, 2012, p 1-14)

Thus, it is a generic representation of a variety of types of applications and implementations within a similar high-level domain. For example, the Open Archival Information System (OAIS) reference model (CCSDS, 2012) powers the community with a high level of abstraction of digital archiving implementation. Furthermore, it is extended to provide a digital archiving service in several domains such as open access repositories (RLG-OCLC, 2002), due to its level of specificity, and thus its ability to consider an open access repository community specific requirements. Hence, its role in the construction of the framework is to provide high level concepts on open access repository in terms of implementations, specifications and functions.

Similarly, the conceptual models and frameworks (see Section 3.3.3 and 5.3.1.2) introduce concepts that are related to each other based on empirical studies. These forms of contracts were utilised hereby to address the conceptual composition in open access analytics using open access repository data. In addition, empirical studies contribute to the process by increasing the propositions and relations between the highlighted concepts.

5.3.2 Phase (2): The Conceptual Analysis of the Process of OAA

With the relevant resources scaled down, based on a systematic filtering process, a careful examination of the relevant concepts was carried out. This section clarifies the phases of literature examination in order to identify, consolidate and link the relevant concepts.

5.3.2.1 Step 1: Relaxing the Ontological Dilemma

The examination of the concepts initiated by relaxing the ontological dilemma exists within the research communities in the relevant literature. This includes: the concept of an open access repository and its complexity across a variety of repositories such as institutional repository and subject-based repository (see Section 2.2.1), the concept of data analytics and its link with knowledge discovery, information visualisation and visual analytics (see Section 2.1.2), the concept of open access repository data (see Section 2.2.2), the concept of social machines (see Section 2.3.1) and the concept of web observatories (see Section 2.3.2). As part of this phase, the clarification of essential terminologies and concepts related to this thesis and the process of OAA-OARD-SM conceptual framework development are given. This is due to the fact that the adoption of a particular perspective or definition that corresponds to the uses of these terminologies is essential.

5.3.2.2 Step 2: Identify the Concepts in the Process of OAA

Based on the definition of the 'process' concept in Section 5.3.1.2 and conceptual synthesis in Section 2.1.3, the process of open access analytics is referred to as the set of activities required in order to develop actionable insights relevant to the problem under investigation. This process can vary from one problem to another, although it can be generalised to some extent with the adoption of process models introduced by the data analytics research community. The adoption and consideration of these models power the framework, with the abstract concepts representing a wider scope of purposes and applications. However, while the delivery of the ROAR analytics service implicitly presents a process model that can be adopted, it still represents a single form of analytics.

This developmental process of the OAA-OARD-SM framework takes into consideration a set of process models contributed by three analytics research communities: knowledge discovery and data mining, information visualisation and visual analytics. The rationale of incorporating a set of process models emerged from multiple analytics research communities to include different forms of analytic applications including descriptive, predictive and prescriptive analytics. In addition, each of these models has its insight in the abstraction of data analytics activities. The following sections briefly highlights these process models and their value to the framework development.

A) KDD Process Model

According to Mariscal et al. (2010), most of KDD approaches are based on Fayyad et al. (1996) process model and the CRISP-DM Reference Model (Chapman et al., 2000). Fayyad et al. (1996) KDD process model (See figure 5.2) frames data mining and knowledge discovery as an integral process that extracts new knowledge from raw data. Their process model consists of five transitional functions that manipulate the raw data to produce a pattern interpreted and evaluated to generate new knowledge. In addition to these functions, the process is initiated by data and domain understanding is carried out to define the KDD process.

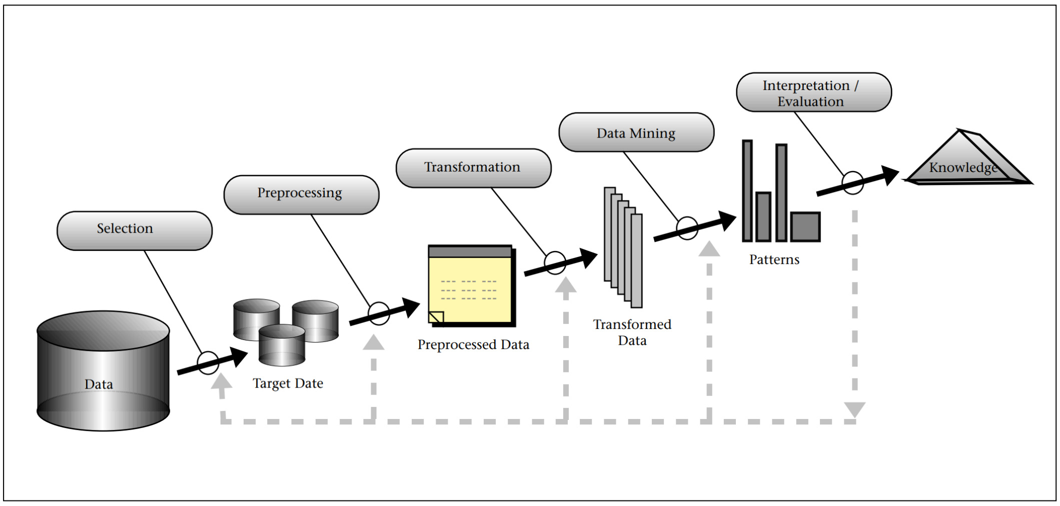


FIGURE 5.2: Fayyad et al. (1996) KDD model and the steps constituting the KDD process.

Similarly, the CRISP-DM reference model (see Figure 5.3) highlights data and domain understanding as the initial process in data mining and consolidates the data selection, preprocessing and transformation into data preparation concept. In contrast to Fayyad et al. (1996) process model, the data mining process is outlined as an iterative process that influences the holistic process instead of being a one-way flow that is associated with a feedback loop during the evaluation and interpretation process. This iteration process demonstrates that an evaluated model can be deployed and maintained over time.

B) Information Visualisations Process Mode

Information visualisation researchers and system designers exploit the power of the reference model to standardise the process of mapping data into visual form (views). Card et al. (1999) reference model interprets the the information visualisation process in the form of adjustable mapping from data to visualisation, where human interactions control the transformation of raw data within three main pipelines: data transformation, visual mapping and view generation.

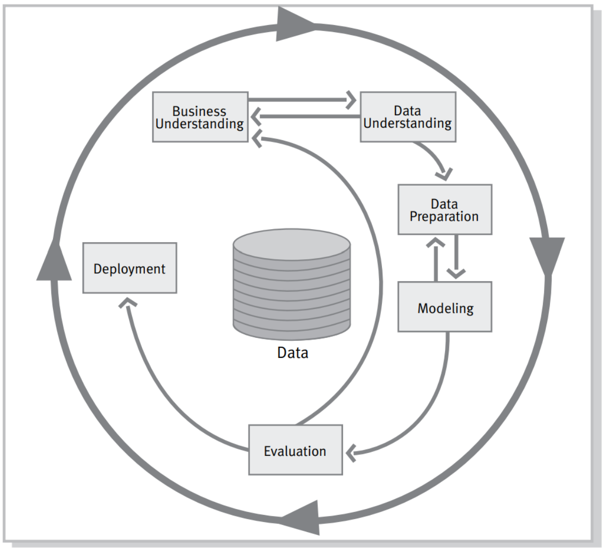


FIGURE 5.3: Cross-Industry Standard Process for Data Mining (CRISP-DM) process model adopted from Chapman et al. (2000)

According to Card et al. (1999) reference model (Figure 5.4), data transformation activities generate analytical abstraction (data tables). Analytical abstraction is a set of attributes obtained from the primary data source and transformed through mapping to visualisation abstraction (visual structure). However, visualised datasets are usually complex and composed of multi-dimensional data at various levels, scales and clusters. Therefore, another process is applied to visual abstraction to transform it to views by users.

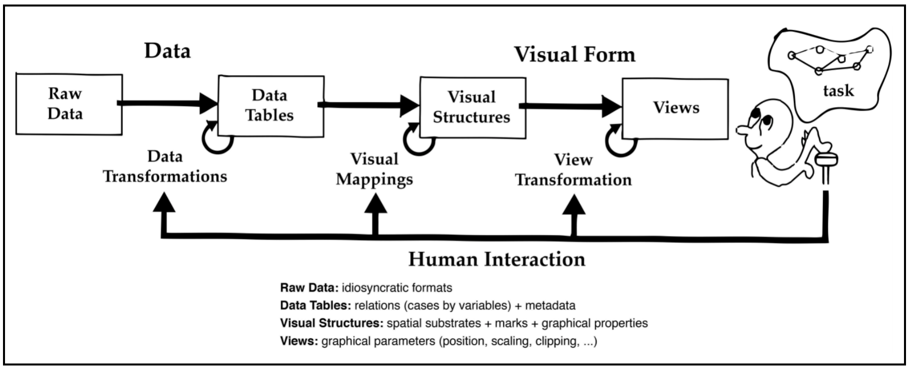

Data
Visual Form
Raw Data
Data Tables
Visual Structures
Views
task
Data Transformations
Visual Mappings
View Transformation
Human Interaction
Raw Data: idiosyncratic formats
Data Tables: relations (cases by variables) + metadata
Visual Structures: spatial substrates + marks + graphical properties
Views: graphical parameters (position, scaling, clipping, ...)

FIGURE 5.4: Information Visualisation Reference Model adopted from Card et al. (1999)

C) Visual Analytics Process Model

With a visual analytics perspective of information visualisation, Wijk (2005) simplified the visual analytics process by dividing it into three processes: visualisation, perception and exploration (see Figure 5.5). Visualisation is central to the analytic process, where the analytics requirements are carried out as a set of specifications that contribute to the visualisation process. Consequently, a set of images (views) are generated that can be perceived by the analyst. The analyst incorporates his/her perceptions, cognitive capability, current knowledge and opinions to generate new knowledge to establish an exploration process with new specification and consequently a new analytics process.

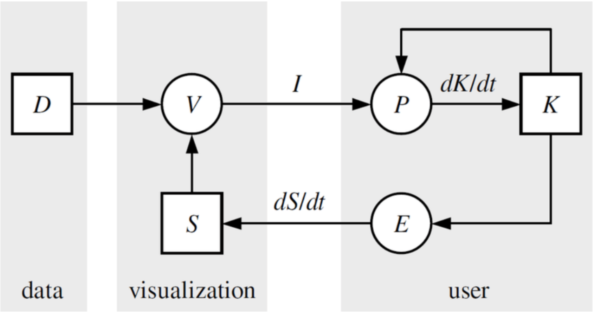


FIGURE 5.5: Wijk (2005) visual analytics simplified process model.

In contrast to Wijk (2005) visual analytics model, which demonstrates the process from a visualisation perspective, Keim et al. (2008) proposed a process model (Figure 5.6) that incorporates both data mining and information visualisation as trajectories for the visual analytics process, a direct data visualisation of the information visualisation process, or the generation of hypotheses with automated algorithms (un-validated pattern). Keim et al. (2008) model encompasses nine processes to generate insight.

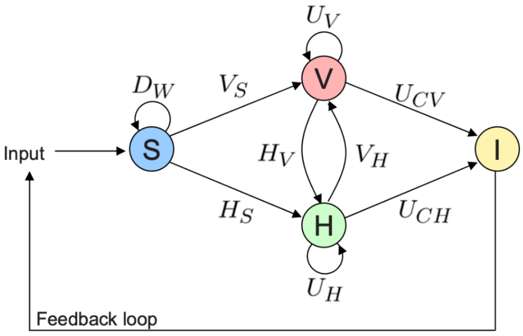


FIGURE 5.6: Keim et al. (2008) visual analytics process model

With a comprehensive view of visual analytics, Sacha et al. (2014) integrated multiple process models adopted from KDD, information visualisation, sense-making and visual analytics research communities into a single 'knowledge generation model'. Their model considers three loops beyond the technological support provided by the computer system, namely the exploration loop, the verification loop and the knowledge generation loop (see Figure 5.7).

D) OAIS Functional Model

The open access repository is proposed to interoperate with the scholarly communication system to overcome preservation issues and provide archiving and preservation services. In addition, the open access repository is implemented with compliance to the OAIS

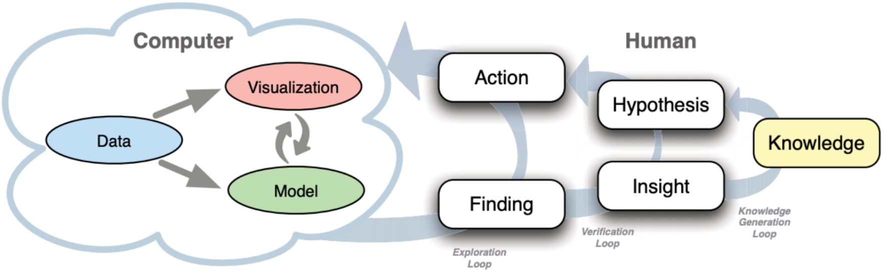


FIGURE 5.7: Sacha et al. (2014) knowledge generation model for visual analytics.

reference model. The OAIS functional model is part of OAIS reference model that encompasses the digital preservation processes into seven functional entities (Figure 5.8), specifically: ingest, data management, archival storage, preservation planning, administration and access (CCSDS, 2012).

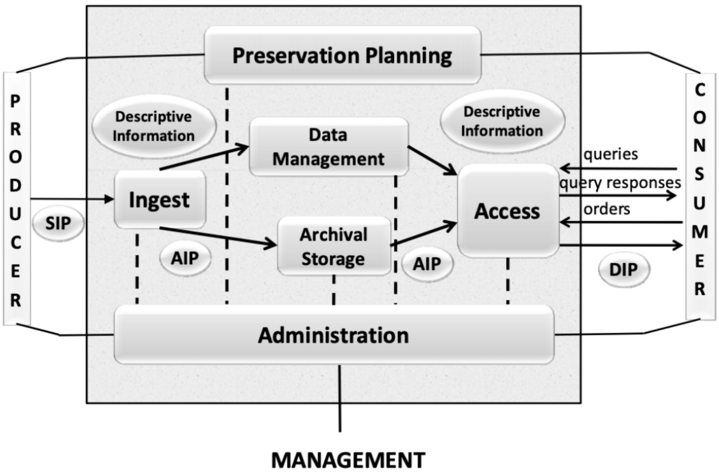


FIGURE 5.8: OAIS Functional Model adopted from CCSDS (2012).

Based on the evaluation efforts carried out by the open access repository research community to refine it taking into consideration the open access repository ecology, a new functional entity was added. A pre-ingest functional entity considers the requirements raised by the complexity of preserving scholarly documents, where the interactions between the producer of preserved resources and the repository need to be highlighted (Allinson, 2006; Nicholson and Dobreva, 2009).

E) Discussion

The KDD process models demonstrate the wide application of data mining techniques to perform the predictive and prescriptive analytics. However, the process of data analytics is automated using data mining and machine learning algorithms (Mariscal et al., 2010). Furthermore, the majority of the KDD process models are data-centric, in contrast to the human-centric visual analytics and information visualisation models (Keim et al., 2008). Thus, the analytics activities are conceptualised as a set of fixed sequences that transform and manipulate the data from one status to another, in order to generate the ultimate output of the KDD process (knowledge).

However, information visualisation incorporates human interactions with the computer-based processing activities. Thus, the information visualisation process is carried out by a set of transformation activities driven by human interactions. The relevance of information visualisations to data analytics is limited to the generation of views from the raw data to support descriptive analytics or as visualisation convey the output of predictive and prescriptive analytics. Yet, human reasoning is out of scope of information visualisation process.

The visual analytics process model integrates human interaction, information visualisation and human reasoning during the analytics process. Thus, its aim is to generate insight instead of views from the raw data. Also, its conceptualisation includes both the utilisation of information visualisation and data mining algorithms during the analytics process. Hence, its conceptualisation is inclusive of all forms of data analytics (descriptive, predictive and prescriptive).

Sacha et al. (2014) knowledge generation model is the model with the most comprehensive and integrated view of the process of data analytics. However, it is influenced by the paradigmatic perspective of visual analytics research on the data analysis process, where the human is central to the analytics process. In addition, the following mantra highly influences visual analytics (Keim et al., 2008, p 164): “Analyse First - Show the importance - Zoom, filter and analyse further - details on Demand”.

Consequently, this places the analytics goal and problem understanding as marginal concept, whereas, the analytics process starts with a direct engagement with analytics process. This is in contrast to the KDD process model, where the goal or problem is central to the process (Mariscal et al., 2010). Thus, within the KDD process model, the process is initiated by problem definition, goal determination and domain understanding, in contrast to the knowledge generation model, in which the goals and aims are emerge by exploration loop and interaction with visualisation and modelling activities.

Wijk (2005) introduced concept specification to convey the requirements of analytic tasks. Also, he distinguished between initial specifications and current specifications. Initial specifications are the specifications that initiate the analytics process and are brought out based on the analyst’s current knowledge, whilst current specifications are integrated into the process over time by an exploration of the generated knowledge.

While the IV, KDD and VA models conceptualise the analytics process, the data collection process and the long-term preservation activities are out of the scope of these models. The open access community harnesses the preservation capabilities of the open access repository to provide free access to open access research. However, its existence in the scholarly publishing system place a layer that collects the research output made open access in distributed nodes of open access repositories. Therefore, in the cases when open access repositories comply with the OAIS reference model, the OAIS functional model abstracts the data collection and preservation processes.

To reduce the complexity of the process model, the OAIS functional entities are consolidated into a digital preservation concept, except for the access functionality, due to the fact its role in the process is as a gateway between data analytics and data collection. This consolidation corresponds with Jantz and Giarlo (2005) definition of digital preservation. In addition, the ingest is scaled down into a self-archiving concept, as the concept is widely used in the open access community to denote the primary approach used to obtain content in open access repositories. The pre-ingest concept is presented as an umbrella that facilitates the process of data analytics, using open access repository data denoted as 'open access reuse'.

With regard to the ROAR case study, analytics services such as ROAR analytics harness the interoperable protocols adopted by the community, whereas interoperability is achieved by means of metadata harvesting. Thus, a harvesting process is part of the data analytics process. In addition, data analytics of large-scale repositories can be performed, and thus, a single point of discovery of these repositories is required. Therefore, repository service discovery is the initial process required before conducting the analytics process of multiple repositories.

5.3.2.3 Step 3: Consolidate and Link the Concepts of the Process of OAA

Based on the above discussion, the following is taking place to raise the processes of open access analytics using open access repository data.

- A set of concepts are adopted from the Keim et al. (2008) and Sacha et al. (2014) models including 'visualisation', 'modeling', 'exploration' and 'verification', whereas, in particular analytics process the 'insight' is determined as the ultimate aim of the analytics process (see the grey coloured parts on Figure 5.9).

- The concept of the social machine promotes a distributed collaboration between the social machine participants in the realisation of a particular process, and it has 'a telos' , that is to say, a general or specific goal (Shadbolt et al., 2019). Thus, the 'Goal determination/Problem Definition' is identified as the concepts that encompass the domain and data understanding, based on the KDD process models (see the blue coloured parts on Figure 5.9).

- The data selection, data preparation and data transformation concepts were consolidated by the data wrangling concept (see the blue coloured parts on Figure 5.9).

- The specification concept is adopted from Wijk (2005) model, in which the initial specification is derived from the analytics goal and the problem definition process, and guided by the data wrangling process, as well as the visualisation and modelling process. However, the current specification is a set of specifications than can be met by the existing data and tools (ex: generating new views) (see the red coloured parts on Figure 5.9).

- To reduce the complexity of the process model, the data collection process within the data provider is consolidated into three main concepts: self-archiving, which represents the ingest of eprints and meta-data; digital preservation, which consolidates a composition of activities to deliver long-term availability of the ingested materials; and the access concept, which abstracts the functions provided by the open access repository to facilitate the use and reuse of their data. Yet, the repository management interaction is considered part of the digital preservation process (see the black coloured parts on Figure 5.9).

- Based on ROAR analytics service, the harvesting and repository discovery concepts are included in the process (see the green coloured parts on Figure 5.9).

5.3.2.4 Step 4: Identify Contextual Concepts Influencing the OAA process

By re-conceptualising the process of open access analytics as an overreaching process that encompass the activities taking place at the repository level to collect and preserve the open access research, the concepts and attributes influencing these processes need to be clearly identified. This can reflect the fact that the context specific concepts need to be highlighted as concepts that enable and constrain the analytic exploitation of the analytic process. Thus, a set of concepts are identified and linked to the three stages: self-archiving, digital preservation and access.

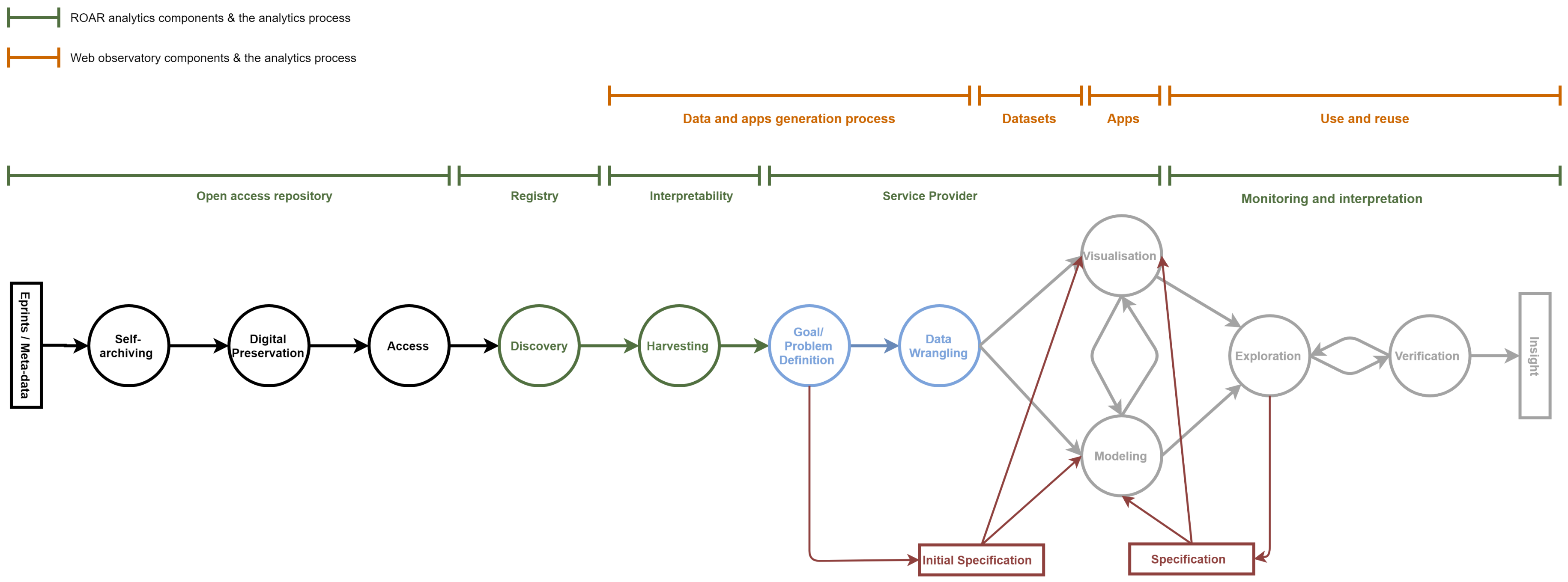


FIGURE 5.9: Synthesised open access analytics process using open access repository data and web observatory.

5.3.3 Phase (3): Conceptual Synthesis of OAA as a Social Machine

The process of open access analytics needs to be refined as a social machine process. The social machine process is characterised by the joints machinery relation of its activities by the human and machine elements within a web-based socio-technical system.

5.3.3.1 Step 1: Visiting the Concept and Implementations of WO

The first step to synthesis the process is the reviewing of the web observatory concept and implementation as the instrument provides the technological components that can support and enable the process of open access analytics. A web observatory is one of the most instrumental technologies that enables collaborators to analyse web activities, and it also enables joint collaboration on the collection and development of tools and applications that power its participants with analytical capabilities.

The concept of the web observatory and its implementation promote a web-based platform that enables datasets, and analytics apps sharing with access control functions underpin it is users. However, its focus is on the use and reuse of contributed data and analytic apps (Madaan et al., 2016). However, one of its strategic aims is to harmonise these analytics and establish collaboration within a particular community (ex web science research community, the health community, astronomy community, etc.). Thus, the web observatory as a concept and its implementation (based on its architectural principles provided in Section 2.3.2.2) are reviewed with the purpose of developing an understanding of how the web observatory as an infrastructure and platform can support the process of open access analytics. Furthermore, the web observatory is envisioned as a coordination platform that enables collaboration within the open access repository community to support open access analytics. Figure (5.9) illustrates the processes fulfilled within the web observatory system by means of using and reusing web observatory apps and tools (the orange coloured lines).

5.3.3.2 Step 2: Identify OAR Stakeholders and their Participatory Role in OAA

The organisational management research community defines stakeholders as a group or individual who can affect or is affected by the achievement of the organisation (Freeman, 1984). In this context, the institutional repository stakeholders are the actors who play an active role in the repository establishment and service delivery or are influenced by their existence. Scott (2009) identified four major group of stakeholders in the information environment in general: end users of information, information providers, information mediators and meta-information users.

A) Information End Users

In Scott (2009) classification, information end users are comprised of a group or individual external stakeholders who need to access information to support their research and work. He highlighted the heterogeneity of this group and their requirements, and the need to be underpinned with tools to identify, locate and access information. Furthermore, the open access community draws attention to the importance of understanding how this layer of stakeholders interacts with specific open access repositories and open access literature in general. Zuccala et al. (2008) emphasised the importance of advocates of the open access repositories, as well as the statistical analysis using download and page linking to understand how effectively the repositories disseminate their contents and are used by end users.

B) Information Providers

In the institutional repository context, information providers are individuals who influence the process of supplying repositories with contents. Scott (2009) listed four contributors to the institutional repository service delivery: authors, publishers, library information services and peer-reviewers. While the first three are very frequently associated with the report context, the peer-reviewer role does not directly influence the repository service delivery and takes place in the publishers' boundaries.

Authors and academics are fundamental stakeholders of open access repositories, as they are primarily designed to collect their research output. According to the Repository Support Project [1], the institutional repositories provide authors with a professionally curated preservation and dissemination environment. In addition to the policies adopted by their institution and funders, they are motivated by the wider access and higher citation rate (Swan, 2010) of their work. However, their engagement is impacted by publisher policies.

Libraries and information services are the conventional managers of repository information resources and collections. With their skills and knowledge about the scholarly publishing system, as well as collection management and intellectual property right issues, they have the opportunity to claim the repository manager role (Walters, 2007). The institutional repository body is becoming more widely appreciated and recognised for their role in their institution, as the repository management profession is acknowledged more by the open access repository community. Wickham (2010) identified three primary roles associated with repository service delivery: repository management, repository technical support and repository administrator.

The repository manager is the professional who manages the human aspects of the repository environment, including (but not limited to) finance management, advocacy, and

[1] The Repositories Support Project (www.rsp.ac.uk) was JISC-funded initiative contributing to building repository capacity, knowledge and skills within UK higher education institutions. It provides the open access repositories community with a set of resources on technical, organisational, management and advocacy sides of the repositories.

external and internal liaison. Zuccala et al. (2008) argued that repository managers should recognise the repository end users, be aware of advocacy benefits and be knowledgeable about copyright infringement and intellectual property issues. In addition, they examined the role of repository managers in five well-known repositories. They concluded that there is a need for ongoing evaluation of repositories management roles to identify the skills and training that support their missions.

According to Wickham (2010), the administration and IT technical support is comprised of software and hardware support, the ongoing administration of repository contents and metadata related issues. They also recognised copyright compliance issues as administrative roles, in contrast to Zuccala et al. (2008) perspective.

C) Meta-Information Users

Both repository managers and administrators recruited by the institution are considered meta-information users. According to Scott (2009), the institution is also a meta-information user stakeholder in the cases that they are also research funders and nation-wide organisations. Typically, institutional repositories are financially supported by an institution to capture their intellectual output and increase the institutional visibility (Kim, 2007). However, research funders influence the author's interaction with repositories by adopting open access policies (Callicott et al., 2015). In the UK, the institutional repositories are given nation-wide attention. The Joint Information Systems Committee (JISC) is a nationwide non-profitable company that supports the digitalisation shift in the UK. It is funded by a remarkable project to support the advancement of open access repositories such as the SHERPA project [2], Repository Support Project and COnnecting REpositories (CORE) project [3].

D) Information Mediator

An information mediator is another external stakeholder of institutional repositories and takes the form of individual or organisation. They provide added value services on top of open access infrastructure. According to Scott (2009), this category is composed of aggregators and search engines.

An example of aggregators is the global virtual archives such as Base (BASE, n.d) and OAIStar (Loesch, 2010), which support their community with cross-repository search functionalities. Another important type of aggregator targets meta-information user stakeholders by supporting them with added value services in an analytical theme (Knoth and Zdrahal, 2012). Therefore, they are typically funded by meta-information users such as the CORE project.

[2] SHERPA (v2.sherpa.ac.uk) is a JISC open access project supports both authors and repositories with tools and services to adopt and compliance to open access polices.

[3] CORE (core.ac.uk)is a service provider operated by the Knowledge Media Institute at Open University. It aggregates open access eprints distributed across different repositories worldwide.

5.3.3.3 Discussion

Base on the overall view of open access repository stakeholders provided in Table 5.4, there two main entities presents with direct interaction and can participate in the open access analytics process; namely, the open access repository management team and the open access aggregator management team. On one hand, the aggregators are highly correlated with the OAI-PMH service provider model discussed in Chapter 4, which may lead to a similar limitation existent in the conventional centralised model. On the other hand, the open access repository has higher access and permission as they are owner of the data or sought a permeation from the original data owner. Therefore, the open access repository manager is determined as a key participant in operating open access. Thus, the process of open access analytics emerges from phase 2 is synthesised as a social machine with consideration of the Smart et al. (2014) characteristics of social machines (see Section 2.3.1), the web observatory infrastructure and concept and the open access repository management team as participants. In addition, a number of open access analytics agenda that can be realised by the open access repository data are determined.

Group	Stakeholders	Role Definition	POAAP[a]	Reference
Information Provider	Repository Manager	A Strategic and financial management, advocacy and communication, staff and project management, expert advice to the institution	Direct	(Wickham, 2010)
	Administrator	Administrative role responsible for managing the content of the repository with respect to data quality and integrity of records, metadata and copyrights. This role usually taken by an information experts and carried out with a close collaboration with the researchers.	Direct	(Wickham, 2010; Amorim et al., 2016)
	Technician	A role with direct focus to technological system implementation and management. They are normally expert in software platforms and the main repository software, deployment, testing, upgrading and development of software instead of the content of the repository.	Direct	(Wickham, 2010)
	Operational Reviewer	Unique role in The University of Surrey repository take place as working group composed of academics, research support staff from all Faculties and librarians that assess the process of the repository meet the Higher Education Funding Council for England (HEFCE) requirements and identify the factors that might increase the risk of non-compliance.	Direct	(Daoutis and Rodriguez-Marquez, 2018)
	Librarians	An existing role in the scholarly communication system where their roles overlap with the role of open access repositories. Therefore, they are presented as the potential stakeholder to take the repository manager role and repository administrator role. Therefore, some repositories are managed by the Librarians and others associated with dedicated departments and management teams. The Librarians play a significant role in scholarly communication and open access publishing advocacy.	Direct	(Schmidt et al., 2018; Daoutis and Rodriguez-Marquez, 2018; Horwood et al., 2004)
Information Mediator	Data Aggregator	A role take place as a form of service provider aggregates the metadata offered by distributed data providers in order to establish added value services on top of open access repositories contents using the standardisation adopted by the open access repositories community. The are an information mediators who enrich the open access repositories data to provides services for end users. They can be general or domain-specific aggregator.	Direct	(Scott, 2009; Knoth and Zdrahal, 2012; Amorim et al., 2016)

Table 5.4: The open access repository stakeholders and the status of their participation in open access analytics process.

[a] The status of their participation in open access analytics process.

Group	Stakeholders	Role Definition	POAAP[a]	Reference
	Value-added services developers	External stakeholders to Open Access Repositories and internal stakeholders to Aggregators utilises the interoperability infrastructure to construct added value services on top of Open Access Repositories Data. They are forced to follow the standard provided by the data providers (Open Access Repositories). According to Houssos et al. (2011), they face challenges to construct custom-made applications that do not follow state-of-t- art interoperability standards. Developers are concerned with the underlying technologies, and in having extensive APIs to promote integration with other tools.	Direct	(Houssos et al., 2011; Meschenmoser et al., 2016)
	Metrics Providers	The operator of specific added value services focuses on analytics and evaluation process through the use of open access repositories data.	Direct	(OBrien et al., 2017; Kelly et al., 2012)
	Information services Providers	The typical added value services providers on the Open Access Repositories Literature such as search functionalities. This type of stakeholders can be internal through individual institutional repository or external through aggregators and providing cross-repositories services	None	(Scott, 2009)
End Users	Researchers as reader	The researchers who utilises the Open Access Resources to conduct their research. They uses the added value functionalities in addition to the analytics tools.	None	(Björk, 2017; Scott, 2009)
	The general public	This types of stakeholder is provided by Björk (2017) with evidence about the advantages of making research open access on the medical sectors and awareness of the state of the art in the scientific research.	None	(Björk, 2017)
	Metrics users	Stakeholders emerge from the analytics dimension that is added to open access repositories data by the scholarly communication community. They include (not limited to) researchers, funders and institution.	None	(Marsh, 2015)
	Research funder	The research funding agencies in general whom invest on research and need to be informed about the impact of their funds and research they are investing in. They have the power to enforce policies (such as mandating open access policies). In addition, they are a major user o f analytics services provided on top of open access repositories data. Example: National Institutes of Health (USA) and Research Councils (UK).	Indirect	

TABLE 5.5: The open access repository stakeholders and the status of their participation in open access analytics process (continued).

[a] The status of their participation in open access analytics process.

Group	Stakeholders	Role Definition	POAAP[a]	Reference
	Nationwide research infrastructure funder	A nationwide organisation and communities support research infrastructure and open access repositories projects. This stakeholders concerned of advocacy, awareness and support of open access repositories infrastructure projects in nationwide scale. Therefore, they have the power to enforce recommendations and policies. The considered as meta-information users or analytics services users as they apply data-based decision model on their funding. Ex: Joint Information System Committee (JISC) in the UK.	Indirect	(UKRI, 2019)
	Regional research infrastructure funder	A regional organisation interoperates the efforts between countries such as EU area. They to some level similar to nationwide research infrastructure funder, however, they aim wider scale. They fund International projects concerns to carry research on open science and open success infrastructure projects provides recommendations for open science community. Ex: EU Example projects: OpenAIRE, OpenAIREplus, FOSTER and FOSTERplus	Indirect	(Bardi et al., 2015; Lossau and Peters, 2008)
	International open access repositories community	International federated organisation provides recommendation and research on open access repositories. Ex: The Confederation of Open Access Repositories (COAR)	Indirect	(Lossau and Peters, 2008)
	Scholarly Publishers	The academics publishers and journals which represents the conventional quality assurance and dissemination entity in the scholarly publishing system. While they are heterogeneous groups of publishers vary in their procedures and policies that influence the way open access repositories operate such as compliance to embargoes periods or copyright transfer agreements placed by group of publishers on their collection. Also, there are models and case studies provided where a strong linkage is provided between open access repositories and publishers.	Negative	(Pinfield, 2009; Schmidt et al., 2018).
	Universities and Institutes	The entity that the open access repository operate under. It is the organisation that the repository operate under and cover the repository operational cost. On the other hand, the repository acts as a central repository for the institution's intellectual output. What motivates institution is the need of recognition and preservation of their research according to the funding institutions requirements. Thus, the institutions value metadata in compliance to standards.	Indirect	(Foster and Gibbons, 2005; Kim, 2007; Amorim et al., 2015)
	Academics as authors	The author who write the scientific materials and utilises the repository to disseminate and self-archive their research. It is the original owner of the open access material which grant it to the publisher and institution though copy transfer agreement or licence agreement. The are the main contributor to the process as they provide their research as open access through self-archiving.	Direct	(Pinfield, 2005b,a)

TABLE 5.6: The open access repository stakeholders and the status of their participation in open access analytics process (continued).

[a] The status of their participation in open access analytics process.

5.4 The OAA-OARD-SM Conceptual Framework

In addition to the graphic illustration of the framework concepts and their relationships, this section provides a discussion regarding open access analytics that use open access repository data with a social machine and narratives on the OAA-OARD-SM conceptual framework. The framework is composed of four layers, presented to show the hierarchy of the structure. This framework provides an understanding of open access analytics as a social machine, where, the distributed open access repositories pool the open access research by the incorporation of authors self-archiving. Consequently, the open access registry enables the collective actions to emerge by wrapping the data sources (the open access repositories) with the web observatory as a coordination platform, thus enabling the data analytics process and achieving the open access analytics insight. The overall framework is illustrated in Figure 5.10.

5.4.1 Layer (1): Open Access Reposioty

An open access repository is a set of systems and services that provide the scholarly community with operational services to manage, retrieve, display and reuse open access research (Pinfield, 2009). While Lynch (2003) presented these services as a layer that are responsible for exploiting and nurturing innovation in scholarly communication, by overcoming the limitations in research dissemination and preservation, this framework positions an open access repository as a layer pool the open access research and enables open access analytics. Positioning the open access repository as a layer with the aforementioned role is inspired and motivated by the efforts made by the web science research community to exploit the analytical value of web data in what is denoted as social observatories (Smart et al., 2019). Thus, in addition to the role of making research open access, this framework emphasis on the realisation of the analytical value of open access repositories by their communities (Carr and Brody, 2007). Although they are not primarily a web origin phenomena, the web science research community have exploited web data in the analysis and insight gathering about social, economic and political phenomena (Hall and Tiropanis, 2012; O'Hara, 2013).In contrast, open access publishing is a web-related phenomenon that has emerged as a result of its existence, where, the open access advocacy community has introduced the web as a means to achieve no barrier access to research output. In this way, research output is considered open access research if, and only if, it is freely accessible through the web (see Section 2.1.1.2).

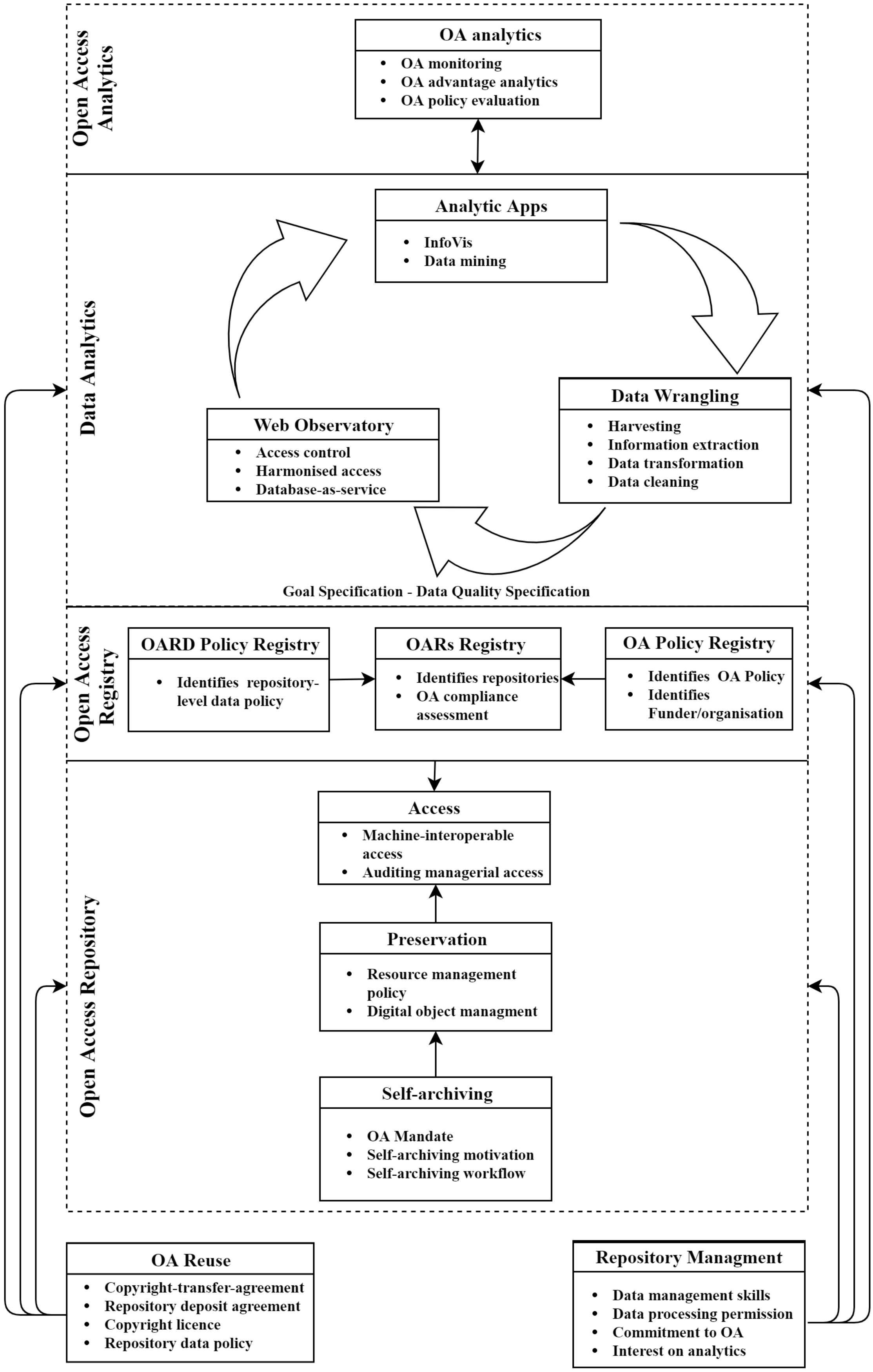


FIGURE 5.10: Open Access Analytics using Open Access Repository Data with Social Machines (OAA-OARD-SM) conceptual framework

.

The open access research community has harnessed open access repositories to realise their goal of open access vision as the primary strategy. It enables them to nurture the author collective deposit action of their research output and ensure free access and the long-term availability of open access research. Yet, the green route is becoming the route with the greatest opportunity of making research open access, as it absorbs the open access research made available on the web (see Section 2.2.3). Also, the Academic institutions, libraries and research funders are the organisational entities facilitating research, as well as supporting the adoption of open access publishing, since the open access repository holds intrinsic value to this organisational umbrella. The intrinsic value is in the fact that scholarly communication forums and knowledge management systems manage institutional research output by capturing intellectual capital of the institutions (Kim, 2007). Consequently, a centralised repository brings together their output and presents a more efficient way of highlighting the institution's research output (Hey, 2004; Bonilla-Calero, 2008) as well as the degree of adoption of open access publishing.

This position is realised through a set of concepts coined to descries the process of acquisition, archiving and dissemination of open access research including self-archiving, digital preservation and access. In addation, it is influenced by open access reuse and repository managers engagement.

5.4.1.1 Self-Archiving

Self-archiving is a broad term used in electronic depositing when done without any publisher mediator (Crow, 2002). In the open access publishing domain, Harnad (1994a) was the first to use the 'self-archiving' metaphor, when calling for a change in the scholarly publishing system from the subscription-based conventional system to a free on-line full-text availability of peer-reviewed research output. Later, the open access community utilised it to describe the practice of depositing research output in open access repositories (Pinfield, 2005b). Hence, it is an approach that allows repositories to make their contents through the collaboration with their users. Harnad (1994a) defines it as the author initiative to

> deposit their refereed articles in ”eprint” archive at their institution [to be] harvested into global virtual archive making its contents freely searchable and accessible on-line by everyone. (Harnad, 1994a, p 1024)

Furthermore, the self-archiving process enables the repositories to acquire two main components of its content: I) eprints and II) meta-data.

I - Eprints

Self-archiving has been made possible because of the shift towards digitisation of the scholarly publishing system. This has brought about a new type of publishing material, identified as eprints. Eprints are the digital form of peer-reviewed articles, which are typically classified either as preprint or postprint. The two categories represent two main phases of the peer-reviewed journal article life cycle. According to NISO/ALPSP JAV technical group, preprints takes the form of 'an author's original copy' or 'submitted mass under review', and a postprint can be delivered as an 'accepted manuscript copy', 'proofed copy' or 'version of record' (Morgan, 2008). These terminologies are utilised by open access repositories communities to identify journal article versions during the publishing process, as a result of the emphasis on open access principles and recommendations for the immediate deposit of research output (Pinfield, 2005b). Hence, authors tend to deposit multiple versions of their work in their open access repositories. Thus, in the context of open access, it is essential to highlight and adopt a version control framework (Brace, 2008).

Wates and Campbell (2007) investigated the changes made in three types of versions, specifically 'author original copy', 'publisher version' and 'accepted manuscript'. Their study showed significant changes are made to the 'author original version' during the peer-review process, with most of these changes are referencing related changes. However, an 'accepted manuscript' is more similar to the publisher's version, which is due to the fact that the 'accepted manuscript' represents an accepted version that requires some reviewing and editing to follow (Morgan, 2008). While the community investigates the influence of article versions on the use of the deposited articles as a source of knowledge (Brace, 2008), to the best of the author's knowledge, their influence on the use of articles as a data source for analytical purposes has not been demonstrated. However, the qualitative analysis of scholarly data and scholarly databases has drawn special attention to the citation data in the published article (Mingers and Leydesdorff, 2015). This is an important indicator when analysing scholarly publishing phenomena. Hence, Wates and Campbell (2007) findings reflect that the process of producing different versions impacts the citation related data.

II - Meta-data

The meta-data is a collateral component to eprints within the repository. Its main objectives are to describe the eprints, support the management of eprints within the repository and to provide access functionalities (see Section 2.2.2.2). Thus, they fall into three categories: descriptive meta-data, administrative meta-data and preservation meta-data (Koutsomitropoulos et al., 2004).

Meta-data is considered as data in the data analytics process models and it has a value in the analytics process (Sacha et al., 2014). While eprints represent a source that data can be extracted from, meta-data is descriptive data that can feed the analytics process. In addition, it is more relevant to open access analytics, as its captures activities within the repository, whether they are done by authors, the system or the repository managers.

In addition, this framework draws attention to the three concepts that influence the self-archiving practices in open access repositories: A) open access mandate, B) self-archiving motivation and C) self-archiving workflow.

A. Open Access Mandate

To accelerate self-archiving in their repositories, institutions and research founders adopt open access mandating policies to encourage researchers to deposit their work in open access repositories. These mandating polices can be classified into two types; I) deposit mandates which demand the researcher deposit their work and II) permission mandates which require the researcher to grant the institution the right to reproduce and distribute the work (Suber, 2008). According to ROARMAP (a global registry for open access mandating policies), more than 900 organisations had adopted open access mandate policies by the end of 2019 [4]. While the open access community evaluates their effectiveness, the potential of open access repositories to support analytic purposes increases.

B. Self-archiving motivations

In addition to mandating policies, there are motivational factors that influence self-archiving practices. There are a set of factors that motivate or hinder authors from depositing their research in open access repositories. Kim (2007) proposed and tested a conceptual model of motivation factors influencing author self-archiving in the institutional repository, based on the socio-technical network model and social exchange theory. Her model was composed of four categories of factors:

- *Cost factors*: a set of concerns that hinder the authors from self-archiving their research, including copyright concerns and time spent and efforts made by the author to deposit their work.
- *Benefit factors*: intrinsically or extrinsically motivate the authors to self-archive their research.
- *Contextual factors*: a set of dynamic factors can enable a high self-archiving rate or act as obstacles to the self-archiving practices. These include factors such as the self-archiving culture, trust, identification and other actor influence.
- *Individual traits*:the author's rank, age and technical skills.

Among all these factors, preservation and copyright are the major motivational factors (ibid). Thus, this framework emphasises these motivational factors as the most significant factors that can determine how much an author self-archive.

C. Self-archiving Workflow

[4] http://roarmap.eprints.org/

Another important concept to consider related to open access repositories self-archiving is the work-flow adopted within a particular repository. This includes the following:

- *Repository platform built-in work-flow*: For example, the widely adopted repository software DSpace and EPrints are designed with a work-flow that enable editors to review the deposit activities to enforce quality control polices.
- *Repository review process and administrative workflow*: Although the type of work-flow adopted is not necessarily related to the software used to establish the repository instead of the institution policy and practice to manage their collection (Carr and Brody, 2007).
- *Mediated deposit*: Beside the repository platform built-in work-flow of deposit activity, the deposit tools that enable a bulk transfer of content into repositories such as Simple Web-Service Offering Repository Deposit (SWORD) protocol (Allinson et al., 2008). Such a tool can automate and mediate the deposit process. Furthermore, it can be utilised to perform a single deposit to multiple repositories and integrate the contents of repositories (Russell and Day, 2010).

5.4.1.2 Digital Preservation

In this framework, the digital preservation concept is used to fulfil two roles: the long-term availability of the self-archived data and the enrichment of the data collected in the self-archiving process. While the first role is met by the essence of digital preservation, as its role is to ensure long term availability and accessibility of the repository data, the second role can be achieved by a set of techniques that enable digital preservation. In their report "Trusted Digital Repositories: Attributes and Responsibilities", the Research Library Group (RLG) defined digital preservation as:

> The managed activities necessary: 1) For the long term maintenance of a byte stream (including metadata) sufficient to reproduce a suitable facsimile of the original document and 2) For the continued accessibility of the document contents through time and changing technology. (RLG-OCLC, 2002, p 3)

According to Pickton et al. (2010), digital preservation in open access repositories can be ensured by understanding the institutional and stakeholder's preservation needs, introducing the preservation policies, and developing a business model to reduce any financial concerns related to digital preservation. In addition to these high-level actions, the implementation of digital preservation and its technology is essential.

Based on the OAIS model (RLG-OCLC, 2002), the core unit of a digital repository providing preservation services is the 'information package', which encapsulates the necessary data and metadata for preservation purposes. Furthermore, three types of information packages are used in open access repositories; the submission information package, the archival information package and the dissemination information package. Each of these is used to produce another type associated with a set of enhancements made to the information package and enriched with the necessary metadata to fulfil its purposes. Thus, eprints and its metadata is encapsulated as a 'digital object' that arrives into the repository as a submission information package. Then, it is transformed into an archival information package, whereas a dissemination information package is generated from the archival information package on demand. An open access repository achieves the preservation function by encapsulating the necessary metadata to reproduce the information package for future use from its archival information package. Although, in addition to access, the analytics practices demand enrichment on the transformation and changes made to a particular digital object.

The open access repositories adopt a set of techniques to enable trust in the open access repository preservation services (Jantz and Giarlo, 2005). In this framework, three of them are highlighted as core functions, including:

- *Digital Signatures*: The digital signature is a method used to detect any change made to a particular digital object in order to operationalise the integrity of the digital object, which is a guarantee that the digital object has not been changed.
- *Persistent Identifiers*: The present identifier is a method of conveying the referential integrity. Referential integrity is essential to ensure that the reference for the digital object is operable and accessible in the long term.
- *Audit Trails*: An audit trail is a method used to capture any changes made to the digital object for maintenance purposes.

5.4.1.3 Access

Access is a core concept in the OAIS functional model that allows the dissemination of repository content. It is enabled through the dissemination information package that encapsulates the digital objects along with the necessary metadata. This framework utilises this concept to operationalise the repository level data collection process in the open access analytics. Thus, the self-archived and preserved digital object, along with necessary metadata, should be accessible for analytical purposes. For this framework, two types of access are distinguished;

A. Machine-interoperable access

Machine-interoperable access is operated to achieve interoperability across open access repositories. Interoperability powers several bodies to exchange information efficiently allowing new knowledge to be generated (Miller, 2000). In addition, the open access repositories achieve the technical interoperability using OAI-PMH protocol (Lagoze and Van de Sompel, 2003).

OAI-PMH is designed with the discoverability of e-prints resources in mind (Van de Sompel and Nelson, 2015), where, a low barrier approach based on meta-data harvesting model is adopted using various standard XML schemes. However, in OAI-PMH implementation, the digital object is out of the scope of OAI-PMH exchange, although its community provides evidence and improvements to its capability to exchange resources. Van de Sompel et al. (2004) demonstrated the challenges associated with resource harvesting using OAI-PMH and introduced new approaches to accommodate resource-harvesting challenges within OAI-PMH based on 'complex object formats'. Their expressive format in resources description enables OAI-PMH harvesters to locate resources and detect any changes made to both meta-data and resources, allowing incremental harvesting in the OAI-PMH framework. The capability of OAI-PMH to intermediate the data exchange across repositories enables for the automatic collection of open access repository data. Moreover, the heterogeneity of both the digital object formats and the software platform powers the open access repository.

B. Managerial Access

The OAIS model illustrates that data management and access are two independent entities. Furthermore, it shows that the access functional entity coordinates access requests that are executed by the data management functional entity. In addition, managerial access is an access request coordinated by the access functional entity and executed by the data management functional entity. The data management functional entity is responsible for backend data management, including the administration of the database system, queries execution and report generation. Thus, this form of access enables access to rich metadata generated by the self-archiving and preservation process.

5.4.1.4 Open Access Reuse

Open access publishing promotes the use and reuse of e-prints published under its concept. This is expressed clearly in the first declaration statement issued after the initial meeting in Budapest (BOAI, 2002). With reference to open access, the Budapest Open Access Initiative (BOAI) (2002) statement went beyond the typical use of e-prints for the purposes of reading, printing and distributing to more advanced uses, such as full-text mining:

> By 'open access' to the literature, we mean its free availability on the public internet, **permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them**

> **for indexing, pass them as data to software, or use them for any other lawful purpose**, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. (BOAI, 2002)

The statement uses a phrase to describe the action of reusing the e-prints as 'input to software' instead of limiting the reuse to a particular application, such as indexing or the application around literature discovery. Hence, a flexible statement accommodates a variety of uses and applications of open access e-prints. The importance of this statement is that it outlines the core orientation of the open access movement by its leaders and initiatives. Also, it is a statement of open access principles (BOAI10, 2012). In addition to promoting reuse, the statement calls for full removal of technical, financial and legal barriers while retaining the author's right to be acknowledged and the requirement that the work only be used for lawful purposes.

Another open access declaration statement was the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (BDOAKSH) (2003) statement issued by the Max Planck Society in 2003. The Berlin Declaration emphasised the BOAI statement principles and promoted the Internet as a functional instrument to achieve open access on a large scale. Regarding the reuse of e-prints, the declaration considered the sustainability aspect of the rights granted over the work made open access by recommending an 'irrevocable' right to access, use, copy and distribute:

> The author(s) and right holder(s) of such contributions grant(s) to all users a free, irrevocable, worldwide, right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship ..., as well as the right to make small numbers of printed copies for their personal use. (BDOA, 2003)

The sustainability and large-scale adoption of open access concepts is a core mission of the Max Planck Society, and 665 organisations already signed the declaration by the end of 2020 (Max-Planck-Gesellschaft, n.d). This indicates the wide adoption of their recommendations and principles put forward in their statements, enabling the large-scale collection of open access resources, which is sustainable with regards to the rights granted to users to use and reuse for lawful purposes.

On the tenth anniversary of the BOAI meeting, the Open Society Foundations issued new recommendations adopted by open access compatible organisations (BOAI10, 2012). The new guidelines brought a concrete proposal for reuse rights. It recommended the CC BY or the equivalent license as 'the preferable' license to be used to make research work openly accessible and support further reuse practices. In short, the CC BY license allows the sharing and adoption of licensed work, which includes copying, distributing, remixing, transforming and reconstructing work on top of licensed work in any medium and format (Creative Commons, n.d).

While the copyright over research and intellectual property rights is a complex area in open access publishing, this section attempts to introduce a set of concepts that influence the reuse of open access materials from the copyright perspective. This includes liber open access vs gratis open access, licence agreement, copy-transfer agreement, licencing, repository data policy and general data protection regulation (GDPR) framework.

A. Libre vs Gratis

Regarding rights, the open access community conceptualisation distinguishes between open access work with reuse rights and open access work with access rights only. Therefore, e-prints with open access rights are referred to as 'gratis open access', and e-prints with open access and reuse rights are termed 'libre open access'. The gratis open access includes scholarly materials without 'toll-access' price. However, any further use that exceeds the definition of fair use requires permission from the copyright owner. Suber (2012, p 66)) described it as "[a] free of charge but not more free than that. Users must still seek permission to exceed fair use".

Hence, it provides access rights only and restricts reuse to cases in which owners have given permission and fair use that is defined by the nationwide legislation. On the other hand, the libre open access is more compliant with the original open access statement issued in Budapest. The libre open access is free of charge and also free of some copyright and licensing restrictions. Indeed, users have permission to exceed fair use: "libre open access removes price barriers and at least some permission barriers" (Suber, 2012, p 66).

B. Licence Agreement

Despite the explicit conceptualisation and recommendations for open access reuse, the use of e-prints published under the open access concept and deposited in open access repositories for analytics practices is more complicated. It involves several challenges illustrated by the open access research community (Jenkins et al., 2007; Gadd et al., 2004, 2003a,b).

The management practices of repositories may comprise several actions, including the saving, copying and aggregation of full texts or partial processing of copyright materials, as well as building added value services on top of deposited e-prints. These activities are regulated and exceed the limits of fair use as defined by national legislation and, thus, require an open licence to be adopted for the deposited work. However, open access repositories are mostly unable to require a particular licence of the deposited work, and the orientation of open access movements to accommodate all types of licence allows open access publishing (BOAI10, 2012).

Therefore, repositories adopt a licence agreement approach. A licence agreement is a legal contract between the copyright owner and the repository institution. As a result of this contract, repositories obtain the right to perform their management, dissemination and preservation tasks on the deposited materials, as well as ask the depositor to take full responsibility for the copyright status of deposited materials. In addition, it represents the process of obtaining permission from the copyright owner to support open access repositories, manage their collection and offer it as open access work for their communities (Gadd et al., 2004).

Open access repositories obtain permission for the deposit process by asking copyright owners to sign a licence agreement or deposit agreement to support their role as open access material providers. Also, the open access repositories disclose the copyright status of deposited materials in the exposed meta-data. Hence, service providers acknowledge the copyright status of harvested materials considering copyright issues when they construct their services on the top of open access repository infrastructure. It is typical for the copyright owner to be the author, and he/she grants a licence to the repository institution by signing the licence agreement. This process is illustrated in Figure 5.11.

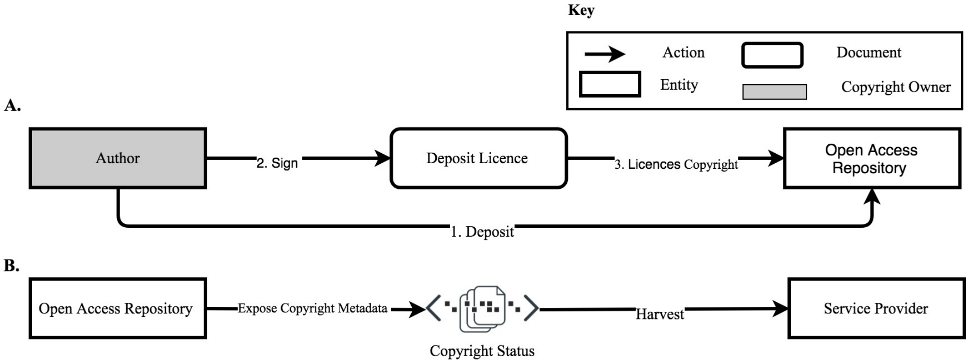


FIGURE 5.11: The permission acquisition process flow within the data service provider framework from the copyright owner (author). A) The repository acquires permission by placing the deposit 'Licence Agreement' within the deposit workflow between the author as the copyright owner and the repository institution. B) The process of disclosing the copyright status of deposit materials to be used by the service provider.

C. Copyright Transfer Agreement

However, in a situation where a copyright transfer agreement (CTA) takes place between the author and the publisher, the CTA is a legal contract between the author and publisher containing provisions to assign full or partial copyright from author to publisher. However, such agreements may conflict with self-archiving practices. Pinfield (2001) showed that academic authors might ignore the terms of their contracts with publishers. Therefore, the licence agreement should include a declaration of the copyright status of the deposited materials and take full responsibility for any copyright infringement action. This approach was emphasised by Jones et al. (2006),

> Such agreement [licence agreement] should ideally be comprehensively gathered at source from the original owner so rights can effectively be passed down the management chain through the institution to the end user with minimal effort. (Jones et al., 2006, p 148)

The flow of a licence agreement is illustrated in Figure 5.12, where the publisher CTA is considered in the process.

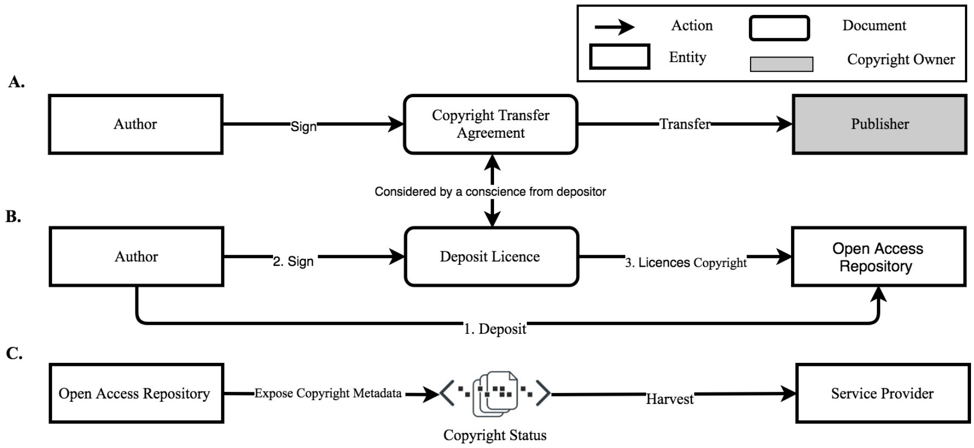


FIGURE 5.12: The permission acquisition process flows within the data service provider framework, taking into consideration the CTA. A) The CTA transfers the copyright status from the author to the publisher. B) The depositor declares all the responsibilities for the copyright status of their work and confirms that no conflict is associated with the self-archiving practice. C) The process of disclosing the copyright status of the deposited work in the exposed meta-data between the data provider and service provider.

Although the CTA varies in the level of right transferred and restriction imposed by the publisher, many publishers recognise the open access policies requirements adopted by the research funders. For instance, some publishers allow the deposit of the pre-print or post-print of their collection in open access repositories (Laakso, 2014). On the other hand, other publishers limit it to particular open access repositories or associate it with research funder requirements (Dodds, 2018).

To highlight this variation and use the opportunity to make research open access, the Sherpa Romeo project aggregates and index publishers' polices (JISC, n.d). Using these policies, 10 pathways are identified to make research open access without infringement of the publisher's policies, including (but not limited to) the pathway associated with prerequisites and conditions, the pathway through APC fees, the embargo period pathway, the licence-based pathway and the publishers' deposit pathway. Although, this pathway highlights the publishers' restrictions and support to open access publishing research instead of their restrictions and support to reuse practices.

Nevertheless, there is a pragmatic shift in the contract-based approach used to regulate the copyright between the author and the publisher. This shift is moving from using copyright assignment using CTA to the licencing approach (Tennan et al., 2019; Dodds, 2018). The open access movement and research funders encourage the authors to adopt the licencing approach, wherein the author grants the publisher a licence to facilitate the publisher with the copyright required to practice their role, while the copyright is retained by the author (Dodds, 2018).

Indeed, there are a set of standards and licences developed by Creative Commons to standardise the licencing of open access work. Creative Commons develops six different licence types, including CC BY, CC BY-SA, CC BY-NC, CC BY-NC-SA CC BY-ND and CCBY-NC-ND. Each of these licences grants a set of reuse permissions to the reuses of licenced work. The most permissive is CC BY licences that allow reuses to destitute, remix, adapt and build upon. Moreover, it allows commercial reuse in contrast to CC BY-NC. All CC BY licences retain the copyright to the creator, and the creator has the right to be acknowledged (Creative Commons, n.d).

With these licences adopted, the conflicts between the open access reuse and publisher practices can be minimised. This is because the copyright is retained on the author's side, who engorged to deposit their work and grant similar licences to the repository. Therefore, this type of licence is attached with the licensed material and exposed in the repository metadata to inform the service provider on the copyright status of the disseminated materials.

D. Repository Data Policy

Although the copyright constringes is not limited to individual e-prints, repositories place data and meta-data policies to protect their collections and highlight the way their collections should be obtained and used. For example, according to the UK-based open access repositories policies indexed in OpenDOAR, most UK institutional repositories limit robotic harvesting to 'transactional' practice, while 'permanent' practice requires obtaining permission from the repository institution.

The open access community constructs registry indexes and disseminates repository data policies, which are used to be exposed in a meta-data format as well. OpenDOAR is a registry that indexes the data reuse policies of open access repositories, allowing service providers to locate the reuse data policy statement and acts based on each repository policies, as illustrated in Figure 5.13.

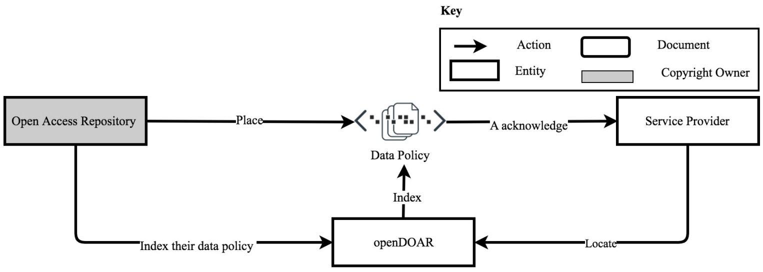


Figure 5.13: The placement of the registry to index repository data policies as an intermediate channel between the repositories and service providers.

E. General Data Protection and Regulation (GDPR)

In addition to the CTA and repository policy, the local and regional legislation can influence the status of the reuse of open access repositories data. Part of this legislation is the GDPR framework that went into effect on 25th May 2018 in EU countries. The GDPR aims to power the data subject with control over their data and inform on how their personal data are being stored, collected and processed. It enforces regulations that cover personal information protection, the right to be forgotten, communication of data breaches, cookies management and consent requirements processes. Therefore, it requires the controller or processor of personal data to adhere to a set of regulations when collecting, storing, processing and sharing personal data about a natural person (Chassang, 2017).

One of the main principles of GDPR is the accountability principle. According to article 5(2) of the GDPR, the data controller or processor of personal data is responsible for compliance to GDPR obligations when collecting, storing, processing and sharing personal data about a natural person. Therefore, it is argued hereby that the GDPR framework can influences the repositories as they are a 'controller' of data that may involve personally identifiable data about the 'subject' from EU countries. According article 4(7) of the GDPR, the controller is "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data".

The open access repositories as a managed entity can take the role and responsibilities of the data controller, as they define their purposes and determine the means of processing. However, this influence depends on the existence of personal data in the open access repositories data. The GDPR states:

> 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier

> or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. 4(1)

The open access repositories store and mange indefinable data such as names, addresses and contacts details either for archiving and administrative purposes. Therefore, they should be compliance to GDPR legislation. The GDPR enforce provisions apply to personal data processing activates and legislates the data processing activities under six lawful basis; the acquisition of a clear consent from the data subject, a possession of contract with the data subject, compliance to legal obligation, data processing for vital interest, performing a public task and processing for legitimate interest (Chassang, 2017).

It also powers the data subject with enhanced personal data protection rights, including the right to be informed, the right of access to their personal data, the right to rectification, the right to erasure, the right to restrict processing, the right to data portability and the right to object and provisions on automated individual decision-making and profiling. Therefore, part of the controller's responsibility is to determine the legal basis they use to process personal data. This includes the purpose of the processing and the means used to process, as well as a clear procedure on how the data will be collected, stored, processed and shredded, in which any further processes can break the purpose limitation principle.

Based on the discussion in Section 2.2.1, the main role of open access repositories is to provide preservation and dissemination services. This role can constitute a legitimate interest between the data owner and the open access repositories, which is the acknowledgement of authorship and the dissemination and preservation of their scholarly work (Phillips and Knoppers, 2019; Smart, 2018a).

Nevertheless, the GDPR gives special consideration to preservation, archival and statistical activities that involve public interest subject to safeguards. The article 5(1)(b) stated:

> Personal data shall be: (b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes.

Besides, in order to compliance with GDPR legislation, the open access repositories adopt privacy notice statement in accordance to articles 12, 13, and 14. This is to establish transparency with the subject on how their data being used and processed.

F. Discussion

Based on the aforementioned complexity of the use and reuse of open access repository data, there are two levels of constraints placed on repository data. First, there are the item-level constraints placed by the CTA or by the national laws that protect the copyright of the author. Second, restrictions also occur at the repository level, in which the open access repository data are constrained by the data policies. In addition, the open access community uses different strategies to resolve copyright issues, beginning with the licence agreement as the open licence of e-prints and the exposure of the copyright status of a particular e-print through a machine-readable format and indexes data policies.

Based on that, the service providers are constrained by the type of licence placed by the copyright owners and should therefore use open access repository data and the copyright status of each item in their mind. However, open access repository management should acquire the type of permission they need to maintain the repository collection through licence agreements, which is designed by them and sought directly from the original owners. Also, they are the owner of the meta-data stored and maintained in the repository. This positions repository managers with the fewest copyright constraints compared with other open access repository data stakeholders.

5.4.1.5 The Repository Managers Engagement

As provided in Section 5.3.3.2, the meta-information users are the primary targets of analytics services provided by the information mediator. The institution as a meta-information user funds institutional repository projects and recruits its management, including repository managers, administrators and technicians. With the issues of reuse influencing the engagement of institutional repository external stakeholders, the institution is the institutional entity proposed by this framework to accommodate the analytics process.

Thus, the repository management operating under the institutional boundaries is expected to interact positively with the framework layers. The engagement of repository management with other repositories, the open access research community and open access advocacy is part of their strategic management responsibilities. According to *"Institutional Repositories: Staff and Skills Set"* SHERPA document, the repository management should have the ability to liaise with repository external stakeholders in the open access movement in general and repository development specific stakeholders, including (but not limited to) funding agencies, publishers, repository groups or federations, service providers, learned societies, international peers and related organisations.

In this framework, the responsibilities of repository managers are extended to include interaction with analytics and web observatories to gather insight from repository data. This interaction is comprised of engagements with open access registries to identify their repositories as open access compliance repositories and engagement with web observatories as mediators in the analytic process. Indeed, engagement with open access registries is well-established in open access communities, with more than 4,000 repositories expected to be indexed in the ROAR registry by the end of 2019 and regarded as the responsibility of repository management team [5].

However, the engagement of repository management team with web observatories has not been demonstrated. The web observatory concept is associated with analytical purposes. Hence, the analytics operations complexity influences their engagement with web observatories. In this framework, the repository management role is extended to accommodate data preparation and transformation responsibility, and establish a data processing unit in order to operate analytics process.

This framework utilises evidence regarding the significance of analytics in terms of repository management and repository management responsibilities, as outlined by the open access repository research community to support this assumption. Zuccala et al. (2008) examined the role of repository managers by surveying five well-known repository management staff. They concluded that the importance of the repository manager to be knowledgeable of tools and methods designed for analytic practices, including citation statistics, download statistics or web link statistics, is essential. In addition, they reflected that it is in the interests of the repository manager to harness the analytical value of collecting university output in a central repository.

Moreover, the SHERPA document outlines repository managers' responsibilities, highlighting the skill sets required to manage an institutional repository, where repository managers need to be aware of the contents of their repositories and any issues surrounding its use for a particular purpose. Also, within the open access community, analytics services are identified as value-added services, whether in reference to individual repository or cross-repository analytics. The cross-repository added-value services are established by the information mediator, and local individual repository value-added services are developed by repository managers. This responsibility is highlighted by the SHERPA document, which states that the repository management should be able to "identify and develop value-added services such as community and collection pages in the repository".

5.4.2 Layer (2): Open Access Registry

The open access repository community establishes a lightweight layer of crowd sourced metadata regarding open access repositories. This layer takes a form of registries index repositories at a single point of access to open access repositories meta-data. These are a collection of records comprised of a set of attributes that catalogue open access

[5] http://rOAR.eprints.org/

repositories. Furthermore, these attributes profile geographical, technical and general details about open access repositories (Carr and Brody, 2007).

5.4.2.1 Single Point of Discovery

The value of open access repositories comes from their value as a whole rather than as a single repository. Thus, the fundamental role of open access registries is to de-fragment the distributed open access repositories infrastructure. Therefore, it is a basic integral component to provide a single global catalogue of open access literature (Harnad, 2006). The open access repository communities actively interact with the registries by identifying new repositories or passively by using the registry index to establish added value services as well as open access infrastructure as a whole, through the use of search engine services (Loesch, 2010) or analytic services (Knoth and Zdrahal, 2012).

To this framework, the open access registries work as web services composition enables the discovery and selection of open access repositories services from a single access point. Thus, they can be utilised to establish a link with other systems and platforms, this underpins the analysis of open access repositories, including web observatory platforms and data wrangling platforms. Accordingly, a fundamental component of this framework enables cross-platform engagement to perform the analysis process.

5.4.2.2 Open Access Concept Compatibility

In addition to their role as web service discovery registry that de-fragments the autonomous repositories nodes forming a distributed infrastructure (Pampel et al., 2013), they act as a gateway for the inclusion and exclusion of open access concept compatible repositories. This role is emphasised in the openDOAR scope of service [6]:

> OpenDOAR has opted to collect and provide information solely on sites that wholly embrace the concept of open access to full-text resources that are of use to academic researchers. Thus sites, where any form of access control prevents immediate access, are not included: likewise, sites that consist of meta-data records only are also declined.

This statement identifies open access compatible repositories through two main principles of open access publishing, specifically the availability of full-text resources and the fact a no access barrier is adopted. On the other hand, ROAR utilises work-flow to review any indexing action made by open access repository managers by an open access concept expert. Thus, the open access registry platform coordinates the efforts made by

[6] https://v2.sherpa.ac.uk/opendoar/information.html

the repository managers to identify their repository with the efforts made by the registry management team to identify the open access compliance repositories.

5.4.2.3 Open Access Policy Aggregation

The open access registry uses a form of policy index to aggregate the open access policies on a global scale. This form of index is significant for open access analytics, as one of its main agendas is to evaluate open access policies on a large scale. In addition to these policies, they identify research funders, institutions and organisations who adopt these policies.

5.4.3 Layer (3): Data Analytics

Generally, analytics is a process comprised of two main phases, including defining problems and the application of analysis methods to gain actionable insight. In addition, the analytics process can take the form of an automated task using data mining a logarithm or take place in human cognition supported by analytics systems and tools (see Section 2.1.2.3). This framework demonstrates it as iterative process that incorporates three main techniques, including data wrangling, web observatories and analytics apps.

Also, defining the problems is the initial step in establishing a particular data analytics process (Cooper et al., 2012). It takes the form of goals identification, domain understanding and determining the data sources to establish the analytics (Fayyad et al., 1996). The availability of repository data without any barriers allows a level of exploration necessary when determining the goals of the analytics process. This initial process enables the generation of analytics specifications that are required to carry out any analytic tasks.

In this framework, each data analytics process is presented as an analytics task, where, each analytics task requires *the analytics goal*, *the analytics specifications* and *the anlaytics apps*.

5.4.3.1 Data Analysis Goal/Problem Definition

Goal determination, or problem definition, initiates both the exploratory data analysis and confirmatory data analysis process. In confirmatory data analysis, it takes the form of a hypothesis (Mulaik, 1985). A hypothesis is defined as a set of assumptions about the subject of the analysis. Thus, it guides the process of data collection, data selection and analysis method determination (Sacha et al., 2014).

In contrast, in exploratory data analysis, the data is collected without a well-defined hypothesis, with the purpose being to formulate one (Mulaik, 1985). Thus, the goal takes the form of the purpose and objective of data analysis determining the problem domain. For example, knowledge discovery and the data mining process is an exploratory form of data analysis where the hypothesis formulation is automated using a data mining algorithm (Fayyad et al., 1996). Hence, the hypothesis is based on unevaluated/uninterpreted data patterns extracted through the execution of a data mining application. In this framework, the goal can take the form of a hypothesis or a general purpose, which guides the data selection process and determines the *data quality specifications*.

5.4.3.2 Data Quality Specifications

Wang and Strong (1996, p 6) defined data quality or data with quality as being the "data that are fit for use by data consumers". The data quality concept is brought into the practice through a set of attributes (data quality attributes) that determine the objective and subjective parameters utilised to evaluate the quality of the data for a particular application (Wang et al., 1993).

Hazen et al. (2014) organised these attributes into four dimensions: accuracy refers to the degree of correctness of data corresponding to the real value, the constancy of data is in its value and structure, the timeliness of the data refers to how up to date the data is, and the degree of data completeness and the fact that it is free from missing data. However, Wang and Strong (1996) and Lee et al. (2002) generalised these attributes into two dimensions: intrinsic and contextualised. While intrinsic refers to the data quality attributes raised by the data itself, the contextualised attributes emerge from the use of data in a particular context.

This data quality dimensions should guide the data wrangling process. Kandel et al. (2011) asserted that the data wrangling process should produce useful data that is usable, credible and relevant to the analysis domain problem. In short, Kandel et al. (2011) research view was based on the generalisation of the data quality attributes into three main dimensions: usability, creditability and relevance. Thus, in this framework, the Wang and Strong (1996) dimensions are divided into these three main dimensions, as illustrated in Figure 5.14.

Within an individual analytic task, what constitutes data quality should be identified and exposed to inform the data wrangling process. Furthermore, the data quality specifications refer to the data quality attributes that should be met by the data wrangling process. In this framework, the data quality specifications are composed of a usability specification, a credibility specification and a relevance specification. The usability specification is characteristic of data used as input for the analytic application. Credibility refers to the trustworthy specification and the characteristics of the data that can meet the analytics goal. Last, relevance specifications are the specifications that determine the analytic task relevant to the data.

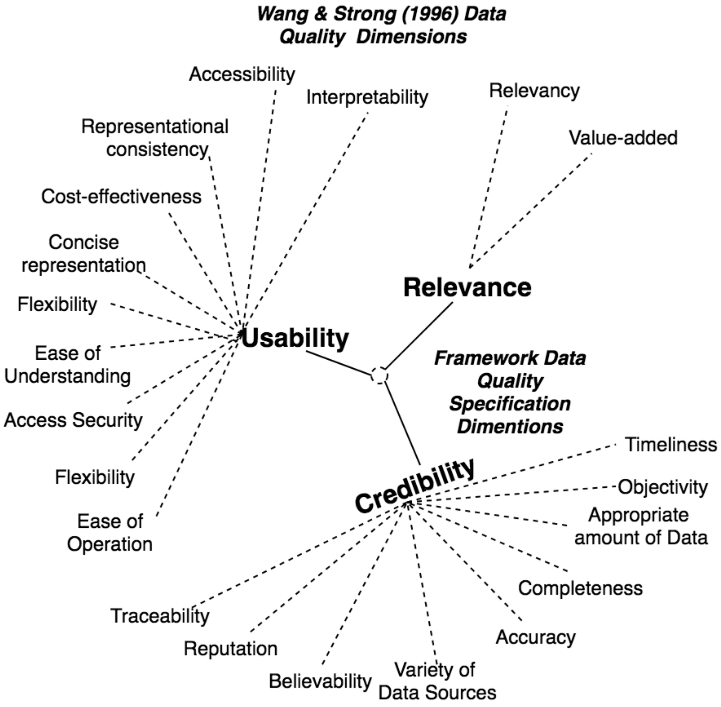


Figure 5.14: The framework of data quality specifications associated with Wang and Strong (1996) data quality dimensions

.

5.4.3.3 Analytics Apps

Analytic apps are a tool that process the output of the data wrangling process to produce visualisations or statistical models derived from a data mining process.

A. Information Visualisation

According to Card et al. (1999) definition of information visualisation and their information visualisation reference model (discussed in Sections 2.1.2.3 and 5.3.2.2), information visualisations involve effectively mapping data and visual representations to support the data sense-making process driven by its user interaction. In this framework, the analytic apps are tools that operate this mapping, generate the visual representation and enable its user interaction.

Federico et al. (2017) reviewed the visual approach to analysing scholarly data. They identified four types of data used in analysis, including text, citation, author and metadata data. These types are associated with three analysis tasks, specifically: element look-up comparison, element relation seeking, synoptic tasks. In addition, there are some approaches that use multiple data types or connect between them, and the analysis task takes the form of aggregation, composition, multiple views or tight integration tasks.

B. Data Mining

The data mining process is part of the knowledge discovery process model utilised to uncover hidden patterns from a large data set. Fayyad et al. (1996) defined data mining as a sub-task of the knowledge discovery process, consisting of applying data analysis and discovery algorithms to produce a particular pattern within the data. The data mining algorithms can be divided into four main types: association rule, clustering, classification and regression algorithms (Hui and Jha, 2000; Nicholson, 2006). Siguenza-Guzman et al. (2015) reviews data mining applications in academic libraries where they highlighted four varieties of applications including service analysis, usage analysis, quality analysis and collection analysis.

5.4.3.4 Data Wrangling

According to Kandel et al. (2011), data wrangling is the iterative process of data exploration and transformation that enables the analysis of data, thus making it useful. The data wrangling process was found to account for 80% of the time and cost in data warehousing projects (Dasu and Johnson, 2003). Concerning open access analytics using open access repository data analytics, the process is associated with high bandwidth and time to harvest the repositories meta-data and resources (Ferros et al., 2008). Thus, in this framework, this task is distributed among repository management, resulting in a collaborative process carried out by a collective approach. Furthermore, this process is informed by the data quality specifications defined in a particular analytic task.

B. Data Harvesting

Despite the challenges associated with OAI-PMH, OAI-PMH metadata presents an opportunity to obtain the open access repository data using a standardised and machine readable approach, where, the availability of open access platforms pave the challenges associated with the harvesting process by providing the community with necessary tools to adopt OAI-PMH standards in distributed archives such as Eprints software and DSpace (Liu et al., 2005). Nevertheless, the OAI-PMH framework can be harnessed to expose richer meta-data schemes such as METS and MPEG-21 DIDL (Bekaert and de Sompel, 2005). In addition, the OAI-PMH achieves interoperability by the means of meta-data harvesting. Thus, the harvesting is the initial step when working with OAI-PMH meta-data. The harvesting process may include both meta-data and resource harvesting, or metadata only, depending on the requirements of a particular analytic task.

C. Information Extraction

Information extraction from the digital object increases its usability for analytics practices. The nature of scholarly documents allows richer information to be extracted and has been utilised for insight gathering. In the context of describing the information extractions process in the CiteSeer project, Williams et al. (2014) addressed six type of information extraction that can be performed on scholarly documents. These include document headers, citations, tables, figures, algorithms and acknowledgement extraction. However, CORE projects limit their extraction to full-text and citations extraction (Pontika et al., 2016). The information extraction process can be simple and limited to the extraction process from meta-data only, but also complex and involving the processing of repository digital objects to extract useful information.

D. Data Errors Cleaning

According to Kim et al. (2003) taxonomical study, there are 33 types of data errors, which are broadly classified into three categories: missing data, incorrect data and inconsistent representation. They concluded that 25 out of 33 types require human intervention. While the technologies supporting data cleaning are advanced, since in Kim et al. (2003) study it was found that human intervention is still essential for the data cleaning process.

E. Data enrichment

Data enrichment is about improving the quality of the data extracted, in order to increase its value for analysis, for example, during the process of correlating data with its meta-data, citation linking and adding a new dimension to the data.

F. Data Transformation

Data transformation involves transforming the extracted and enriched data to unify the scheme to achieve data usability for a particular data analysis task. Kandel et al. (2011) argued that useful data is the data that can be processed and manipulated by analysis tools. The transformation process moves the data from its raw form into feature data or analytical abstract to be processed by analysis tools (the analytics apps).

5.4.3.5 Web Observatory

Web observatories are distributed infrastructures that support resource sharing while privileges to view, query and download are controlled by their owners (Tiropanis et al., 2014a; Tinati et al., 2015). In this framework, the web observatory concept, along with its architectural principles (provided in Section 2.3.2.2) contributed by Tiropanis et al. (2014a,c), are adopted. Its role in the framework is to support the coordination of the analytics process in a distributed manner. In short, its role is to underpin the following functionalities:

A. Database as a Service

The typical approach to data warehousing and decision support systems is to maintain the analytical data in separate data stores from the operational or transactional data

(Chaudhuri and Dayal, 1997). This is due to the variance in functionalities and performance that are required to achieve their objectives. An open access repository is designed to deliver the preservation and dissemination services, thus, their systems are not intended to accommodate analytical functionality. This framework adopts the capabilities associated with allocating cloud databases by the repository managers to host harvested and processed data extracted from repository data, and processing unit process analytic applications queries. To avoid the technical complexity of engagements in repository observatories and to minimise costs (Abadi, 2009), the database as a service prototype is adopted (Hacigumus et al., 2002). The opportunities of the database-as-service prototype for analytical application, in term of supporting heavy update applications and ad-hoc analytic and decision support, were discussed by Agrawal et al. (2011).

Being the databases allocated by repositories managers, the cost of data processing and storage are distributed among institutions with a repository, instead of a centralised organisation. As reported by Greenberg et al. (2008), the high cost of establishing a data centre in the cloud is associated with the services and computational resources allocated to run the data centre functionality, which represents 45% of the total cost. However, that makes the web observatory a trusted zone stores the links to database resources. Therefore, it should securely maintain access to catalogued databases.

B. Access Control

One of the four main architectural principles of web observatory architecture is to support both closed or open licence data. Therefore, it provides a technique that enables data owners to control access to their data. As well as that, data users can obtain access to catalogued data from their original owners. Thus, it is anticipated by the web observatory community that it will bridge the gap between big data and private data (Tiropanis et al., 2014a; Tinati et al., 2015). This principle is significant for open access repository data, due to the copyright concerns associated with the use and reuse of repository data (see Section 5.4.1.4)

C. Harmonised Access

From a technical point of view, the web observatory architecture is designed to accommodate a heterogeneous data format, including (but not limited to) flat file stores, SQL, NoSQL and triple store formats. Therefore, it works as a reverse proxy pipeline the distributed analytics apps with their remote datasets. In turn, this can reduces the technical heterogeneity and coordinates the analytics process with minimum requirements (ibid).

5.4.4 Layer (4): Open Access Analytics

This layer represents the utilisation of the analytics activities as a whole, instead of those of an individual repository, which enables the realisation of the open access analytics agenda through a set of exploration and verification loops directed to gather insight about open access publishing including the open access adoption and coverage analysis, open access policy monitoring and evaluation and open access advantage analysis.

5.4.4.1 The Open Access Adoption and Coverage Analysis

The demand for open access publishing adoption to be monitored is another consequence of the distribution of open access infrastructure and the venues involved in supporting open access publishing. Thus, to understand the progress and its dynamic, distributed sources and publishing venues need to be examined and analysed to deduce the changes taking place in scholarly publishing system (Jubb et al., 2017).

These types of analytics take the form of institution-wide, nation-wide and cross-nation wide practices. A local monitor is required to monitor open access publishing for a particular institution or academic venue. A nation-wide monitor extends its service to cover specific nation institutions, through the coordination of institutions and the analysis of data from publishing venues (JISC, n.d.a). A cross-nations monitoring service is a service analyse the open access publishing in a set of countries which provides a concrete indicator of open access publishing adoption progress (Bardi et al., 2015). Another group takes the form of quantification services, which quantify institutions' deposits to understand the open access repositories dynamic(Carr and Brody, 2007). These type of monitoring services are not necessarily concerned about open access peer-reviewed publishing, but rather the growth of open access repositories. To all these types services, open access repositories are a fundamental data source, as they are a primary green open access research provider.

The monitoring of open access growth and adoption can be evaluated from a repository level, where the analytics subject refers to open access repositories. In addition, the agenda of analytics is to gain insight into the distribution and the growth of the adoption of open access repositories. For example, open access repository registries track open access repositories details, including their geographical data, which is visualised by the Repository66 analytics service [7].

5.4.4.2 Open Access Policy Monitoring and Evaluation

Open access policies are principles of action placed by research funders, institutions or governments to impose mandatory open access publishing of research output or request

[7] http://maps.repository66.org

open access publishing of research output under their grant or umbrella (Harnad et al., 2008). The Open Access Community gives considerable attention to these policies. Thus, a global index has been established [8], and evaluation studies have been carried out to evaluate their adoption and measure their effectiveness (Swan et al., 2015).

Vincent-Lamarre et al. (2016) evaluated the effectiveness of open access policies and found that various types of policies are adopted by different institutions, funders and governments, which in turn enabled them to score the effectiveness of a particular type of policy. Thus, they introduced a guide for institutions to determine an effective framework for their open access policy. Open access repositories and global scholarly databases were the main data sources that contributed to their research.

5.4.4.3 Open Access Advantage Analysis

The open access research community promotes and advocates open access publishing through a set of strategies. One of these strategies is to reveal its positive impact on research in terms of visibility (Ghosh, 2011). Thus, considerable efforts have been made by open access researchers to contribute to the debate of open access citation, regarding its impact and visibility advantages (Swan, 2010). Concrete evidence based on the bibliometric and statistical analysis of open access to non-open access articles has been established (Eysenbach, 2006), by examining this advantage across research fields and in a large scale manner (Harnad and Brody, 2004).

The data used to carry out those studies was obtained from leading scholarly database services such as Scopus and Web of Science. However, the adequacy of the data available for such studies remains one of the challenges to enable robust analysis. Swan (2010) argued that it is challenging to collect a sample of a critical size to study the open access advantage in some fields. Also, it is essential to determine the publishing data, yet the date when the article was made openly accessible is not always stated, and it is challenging to establish.

Open access repository data is one of the source's power these form of analytics, where its data is used to represent the immediate open access activities, which enable it to be examined in comparison with non-open access articles (Brody, 2004; Moed, 2007). In addition, open access repositories also contribute to methods of seeking such evidence. For instance, Brody et al. (2006) used the download/citation correlation in order to predict the citation of scientific articles in its early stages, as this is one of the indicators of access to research. Consequently, it may reflect theuse of particular research, where open access articles downloaded three times as non-open access articles (Harnad, 2005).

[8] https://roarmap.eprints.org

Seeking this form of evidence is one of the analytical agendas of the open access community, and as well as open access repositories data contributing to previous research, they will still contribute to future investigations.

5.5 Chapter Summary

This chapter reports the process of developing the OAA-OARD-SM conceptual framework which discusses the process of open access analytics on top of open access repositories. Besides, it adopts the social machine concept in the realisation of the open access analytics process. Accordingly, the process of open access analytics is conceptualised in four main layers, namely the open access repositories layer pools the research made open access, the open access registries layer wrap, identifies and defragments the open access repositories, the data analytics layer encapsulate the analytics activities within individual repository mediated by the web observatories infrastructure and the open access analytics layer take advances of insight gathered to realise open access analytics agenda. One of the important propositions made by the OAA-OARD-SM conceptual framework is the role of open access repository management team in the realisation of open access analytics.

Chapter 6

Interviews with OAR Management Team Members

6.1 Introduction

The conceptual study reported in Chapter 5 provides understanding on provided an understanding of the open access analytics from a process level, as the process was re-conceptualised with incorporation of the social machine concept, where, the concept of open access analytics is re-framed into OAA-OARD-SM conceptual framework. One of the main characteristics associated with the OAA-OARD-SM framework is the active, participatory role of the repository management in the realisation of the open access analytics process with an individual repository, which can contribute to open access analytics as a whole. Their participatory role is inductively deduced based on the existing research, which highlights their interest in the analytical value of their data. In addition, the studies also highlight the roles of repository management and their responsibilities regarding the repository. However, there is little research into how repository managers harness their repository data for analytics practices and how they interact with analytics systems and services operating on top of their data. Therefore, this chapter reports the findings of a qualitative study which was conducted to investigate the exploitation of open access repository data for analytics practices. Thus, it focuses on the analytics practices at the repository level. Consequently, the research question the study is:

> RQ.3: How is the open access repository data exploited for analytics practices in the UK based on open access repositories by their management?

This chapter goes through the research design procedures, including details related to the interview instrument design, research sampling and participant recruitment. What follows is the interview findings grouped as themes and discussed considering the open access repository literature. Lastly, the chapter provides a further discussion on the OAA-OARD-SM conceptual framework. The study provides insight into how repository data is exploited at the repository level and calls for the re-construction of this exploitation using the OAA-OARD-SM.

6.2 The Study Design

This study utilises a qualitative research design to acquire an understanding of the process involved in the data exploitation of open access repositories for analytics applications in UK-based institutional repositories. The indicative approach used in the qualitative research and its advantages to provides in-depth understanding of a particular phenomenon can support the achievement of the aforementioned study aim (Maxwell and Loomis, 2003). According to Merriam (1988), the aim of qualitative research is to capture the process by which the action takes place, rather than concentrate on the outcomes. Whereas, Furthermore, the strength of qualitative research lays in its ability to uncover the circumstances that lead to the outcomes, compared with quantitative approaches that are directed to capture the outcomes (Maxwell, 2005).

A further issue is the limitation of the existing research to inform a quantitative analysis in the topic of open access repositories exploitation for analytics applications. This is due to the limited research conducted into analytics in the open access repository literature and the fast-changing technical and organisational landscape of open access repositories. Thus, the exploratory and inductive natural of the qualitative research enables the researcher to carry out the research process with minimal assumptions (Patton, 2002; Myers, 2013).

The overall approach used in this study is interviews with experts, which was achieved through the incorporation of semi-structured interviews (see Section 3.3.7) as the data collection method and thematic analysis (see Section 3.3.8) as the data analysis method. The approach taken informs the research design in terms of the selection of participants and their sampling, as well as how the knowledge needs to be acquired and analysed (see Chapter 3). In this case, the semi-structure expert interviews (see Section 3.3.6) were used as a collection technique, enabling the researcher to collect rich, comprehensive and thorough data about participants' experiences regarding the issue under investigation (Creswell and Poth, 2018; Turner III, 2010), with a level of structure imposed by the collection instrument. Accordingly, the thematic analysis makes use of this rich data to provide an understanding of the matter in hand, by highlighting a pattern in the raw data (Braun et al., 2018).

6.2.1 The Study Participants

Open access repositories mostly follow a widely adopted interoperability protocol: the OAI-PMH. This is composed of two basic entities: the data provider and the service provider (see Chapter 4). The open access repository is the data provider entity, and the service provider is represented by the open access aggregator, which harvests the repository metadata to build and operate the added value services for various applications, including the analytics applications. Therefore, it is logical and appropriate to investigate the exploitation of analytics from the side of the service provider.

However, this study seeks to understand this within the repositories and not be restricted to OAI-PMH based analytics applications. Thus, the OAI-PMH based application only forms part of the picture, instead of being the ultimate answer to the research question (RQ3), based on the assumption that their management team are aware off the analytical exploitation of their repository data as reflective subject matter expertise (see Section 3.3.6).

One of the main principles associated with the use of qualitative interviews and the human participant as an instrument to collect data is the correlation between the participants and the issue under investigation. Foddy and Foddy (1994) asserted that the participant should possess and have access the information required by the investigator, so that no barrier may influence the data collection process. This also align with 'competence of experts' strategy adopted in this study to identify experts (see Section 3.3.6). Consequently, any participants included in this research needed to be aware of the open access repository's internal operations and interact with the repository closely.

open access repository ecology involves several stakeholders, including internal and external stakeholders (see Section 5.3.3.2). External stakeholders have insufficient access to the repository operational issues, while the internal stakeholders operate and facilitate the repository to deliver its services to its designated community. Internal stakeholders include repository institutions, funders and management, and while the first two take the role of stewardship of the open access repository, the third takes the role of operating and managing the open access repository service, including analytics services. Thus, they obtain their knowledge from practical experience, and are therefore aware of any changes to the system's operation and application.

However, the position of the participants in relation to the open access repository is not the only aspect that requires prior assessment, as the participants' experience of open access repository management is also critical. In addition, since open access repositories are a relatively new entity in the scholarly communication system, responsibilities regarding their management have often been coordinated with existing roles in the academic institution such as librarians (Ottaviani and Hank, 2009). This was particularly the case before it became apparent that their role is unique, as it focuses on digital collection management (Walters, 2007) and research data management (Swan, 2012). Additionally, repository management roles are often a part-time job or combined with other roles (Wickham, 2010). This may influence the number of years a repository manager commits to the position. Given these issues, one year of experience of managing repositories is determined as a minimal inclusion for participation in the interviews for this research.

6.2.2 The Purposive Sampling

According to Maxwell (2005), the sampling strategy for qualitative research generally falls under a strategy called 'purposeful selection', 'purposeful sampling' or 'criterion-based sampling'. Furthermore, he quoted Weiss (1995) important statement on the nature of participant selection in qualitative research design to emphasise the dynamics of the participant selection process, which varies from one qualitative research to another, as researchers identify their participants for a specific purpose. Weiss (1995) also argued that a researcher using qualitative interviews should not use samples, but instead use a panel of people who are uniquely able to inform the researcher about the subject or event being investigated, since they are an expert in the subject or witness to the event. Lastly, qualitative research generally involves a fewer number of participants or cases compared to quantitative research, yet it provides an in-depth analysis of the subject under study, and thus a deeper understanding of it (Baker and Edwards, 2012).

In purposive sampling, the researcher attempts to target the information-rich cases for the most effective use of resources (Patton, 2002). However, it is not simply sufficient for the case to information-rich, as the case should be available for investigation, and the participant willing and cooperative, in order to enable a productive data collection process. In order to achieve this, an expert in the repository landscape of UK institutions was recruited to identify a set of UK repositories that could be investigated in terms of the analytic exploitation of their repository data. Accordingly, eight UK-based institutional repositories were selected as potential cases to carry out this study.

Another essential concept of qualitative research subject selection and sampling is the *data saturation* concept. As a result of the exploratory nature of qualitative research, the researcher is not aware of how many cases or participants are enough to address the research question. Thus, the researcher examines whether the cases are sufficient or not as part of an ongoing process. The researcher does so by testing the data saturation in the collected data. The data saturation is the degree to which data is expressed, thus repeated what has been shown by the previous data (Ness et al., 2015). Although data saturation is an important concept, it is challenging to practice it. In an expert review carried out on "the number of interviews sufficient to conduct qualitative research" report Alan Bryman argues that using the data saturation strategy 'forces the researcher to combine sampling, data collection, and data analysis, rather than treating them as separate stages in a linear process' (Baker and Edwards, 2012, p 5).

In this research, purposive sampling was used, and the data was examined for data saturation at the end of the study. It is argued here that the study research reached saturation after only a few cases had been examined, due to the nationwide coordination enabled by JISC between the repositories in the UK.

6.2.3 Interview Questions Design

The interview question design was based on the main view of the role of qualitative interview, which is that it is used to collect rich data about the issue at hand. In contrast to a quantitative research design, the relation between the research question and interview question is a systematic conversion through a research question related concept operationalisation process. That is to say, the qualitative interview design is based on the goal of the data collection. Maxwell (2005) stated that:

> Your research questions formulate what you want to understand; your interview questions are what you ask people in order to gain that understanding. The development of good interview question requires creativity and insight, rather than a mechanical conversion of the research questions into an interview guide [...], and depends fundamentally on how the interview questions [...] will actually work in practice. (Maxwell, 2005, p 92)

The interview questions in a qualitative interview are denoted as an 'interview guide', 'interview schedule' or 'interview framework' (Kallio et al., 2016). These terms highlight the high level of control placed by the interview questions during the data collection process. Three main sections were determined; the analytics system and services operated on top of open access repository data, the interaction between the repository management and these systems and services, and the influences and concerns regarding the operation of these services on the repository service delivery. In addition to these three sections, another section was included to collect a set of information about the participants.

Section (0)

The participants validation process: The interview began with validation procedures designed to collect information about the participant's years of experience and their role in the managing of the repository (see Appendix A). Walters (2007) identifies three main roles associated with repository management: the repository administrator, repository manager and repository technician. However, in practice, repository management may vary and be more complex in its structure. Thus, an alternative open question was provided to allow participants to specify their role. The use of an open-ended question to acquire role-related information may have added unnecessary complexity to the research, therefore the open-question was followed by a brief clarification of the role identified, in order to correlate it with the pre-defined role.

Section (1)

Naming the analytics systems and services operating on top of the open access repository data: The interview as a data collection method was based on a common understanding between the interviewee and the participant. This is because the concept of analytics is an emerging concept that is seen from various perspectives that differ from the one defined in this thesis (see Sections 2.1.2 and 2.1.3) As a result, it was deemed possible that this may influence the data collection process. Thus, the researcher provided each participant with a brief definition of data analytics for this research at the beginning of each interview. However, the participants lost sight of the concept of analytics as the interview progressed and moved on to other sections. In order to counteract this, this section used to acquire a set of information about analytics systems and services used or those fed by the open access repository data. According to Sacha et al. (2014) knowledge generation model for visual analytics, the analytics process is carried out along with human and technological input that work together to visualise and model the data. Thus, the identification of these systems and services can serve as a representative of particular analytic applications of open access repository data. Whereas, a set of probes can be made around particular system and services. This also serves as the means for other two purposes: data triangulation using secondary data, and to examine these services, in terms of whether they involve the exploitation of open access repository data and whether they are open access analytics.

Section (2)

Uncovering the interaction between open access repository and its management within a particular analytics system and service: To understand a particular analytic application, this section focused on two main aspects, specifically the type of interactions that exist between open access repositories and analytics services and the participatory role played by the repository management within a particular analytic application. Based on the theoretical discussion of a set of pre-identified open access analytic uses of open access repository data in Section 2.2.3, analytics systems and services can be operated on a variety of scales such as a local scale within a particular repository or institution, a nation-wide scale, a regional scale and/or a global scale. Thus, part of understanding these analytic applications was to capture the coordination and communication process between the repositories and the analytics service providers at the repository level.

Section (3)

The influences and concerns of these services operations on repository service delivery from the perspective of repository manager: With a clear view sought on the interactions between the repository and analytics systems and services, this section was dedicated to understanding how these interactions influence the operation of the open access repository and deliver their main services, which is their dissemination and preservation services. In addition, it sought to gain their perspective on the concerns around the operation of analytics services using open access repository data, as well as their reaction to such concerns. The interview questions were debriefed and reviewed (Creswell and Miller, 2000) by an expert in open access repositories and open access publishing research. The semi structured interview questions used to collect data are provided in Appendix (B).

6.2.4 The Participant Recruitment

In order to recruit the repository management team members as experts, repository web pages were used to collect contact details provided publicly. This reduced the ethical concerns associated with collecting participants' personal contact details. These contact details were used to send an invitation email to the repository management team, asking them to participate in the study, with the value of their participation being highlighted (see Appendix C).

Out of eight invitations, four positive responses were acquired. To find four further participants, each institution website was searched to understand the way their repositories were managed, in terms of which department took responsibility for the repository management, and any department contact details made publicly available were collected. Armed with the new contact details, follow-up emails were sent to the four other repositories, encouraging them to participate in the study. This follow-up process recruited another two repositories, while no response was acquired from the other two. Table 6.1 provides a comprehensive view on the participants recruited in this study.

P(No)	Repository No	Participant Role	Level of Experience
P(01)	R(01)	Technician and User Support	5-10 Years
P(02)	R(01)	Repository Manager	1-5 Years
P(03)	R(02)	Research Information Manager	> 10 Years
P(04)	R(02)	Repository Administrator	> 10 Years
P(05)	R(02)	Repository Manager	> 10 Years
P(06)	R(03)	Repository Manager	> 10 Years
P(07)	R(04)	Repository Manager	5-10 Years
P(08)	R(05)	Repository Manager	5-10 Years
P(09)	R(06)	Repository Administrator	> 10 Years
P(10)	R(06)	Repository Manager	1-5 Years

TABLE 6.1: Study participants' role in repository management and level of experience.

For each particular repository, the participation of its manager was determined as a minimal requirement, and other members of the management team were also encouraged to participate. For example, in R(01), the manager advises the researcher to conduct another interview with the technician and user support manager, as she was well engaged in the development and operation of repository services and the current research information system. Furthermore, in R(02) both the repository manager and the research information manager were also recruited. The opportunity to recruit more than one participant from a particular repository also depended on the approach used to manage the repository and the attention given by the institution to the repository. Indeed, two of the cases were only managed by one person (the repository manager), who liaised with the other departments in the institution (ex: IT department). Although, the ability of other members of the management team (other than the repository manager) was found to be influenced by the same issue. In R(06), a repository administrator was interviewed, and he emphasised his limited interaction and awareness of the analytics application within their repository. P(09) stated that:

> I don't think that I can give too much on that [analytics] ... I don't have much to do with that ... I don't have to in my role. P(09)

He added that:

> a part of what happened with analytics it doesn't relay anything to do with this, if it does, our manager will probably come to us, and say look we found this out, and we need to do this, we need to do that. So they [editorial team] stopped with that we are part of the team. P(09)

However, he conceded that it might be different in other repositories. In this case, their role was to take action instead of performing the analytics or interact with analytics.

> In a way we quite lucky that we have a separate editorial team [...] because I am aware of other repositories you could have somebody have to deal with a quite high level as well as fit that in [...] we have the resources for the team but I am aware of other places they don't, so you could have somebody in my position somewhere else who does have to deal with analytics. P(09)

In contrast to the R(02), the repository administrator was directly engaged with some analytics that is correlated with the administration role or requires their engagement to support its function. Such dynamic in the repository management structure was considered.

6.2.5 Ethical Issues and Ethical Approval

According to the university ethical framework, any research carried out within the university umbrella involving human participants should receive an ethical review. Therefore, this study was ethically approved by the university ethical committee under number 48465. Also, the interviews were audio-recorded. Thus, consent was sought from each participant to record their voices, which included them providing their name and signature. Furthermore, part of the ethical and privacy issue is reporting repositories names or institutions the participants belong to. The institutional repositories management members are few and in some cases, the repository managed by one person. Hence, identifying the repository or the institution can indirectly identify the participant. The repositories and their institution were identified and used for sampling purposes and data translation, although they are not reported in this thesis — also, the services which are uniquely provided within one institution where anonymised.

6.3 Finding and Discussion

This section presents the study findings and discusses them in the light of open access, open access repositories and data analytics research.

6.3.1 Section (1): Analytics Applications within OARs and the Exploitation of OARD

The study participants highlighted nineteen (19) analytics systems, tools and services used by the repository management or fed by open access repository data, although these systems, tools and services were not necessarily established on top of open access repository data and not necessarily carried out for open access agenda. Tables D.1, D.2 and D.3 in Appendix D list them with their status, in terms of whether there was an exploitation of open access repository data and whether it could be considered as open access analytics (see Section 2.1.3).

Further analysis was carried out to investigate the services and tools associated with open access repository data analytical exploitation's, in order to categories them. This analysis was based on the secondary data collected from the system/service/tool webpages or scholarly publication describing them, in addition to the interview data collected on the links between these services and open access repository data.

Based on this analysis, two main categories were identified: the locally operated service/system/tool and the distributed services. The position of a particular analytics

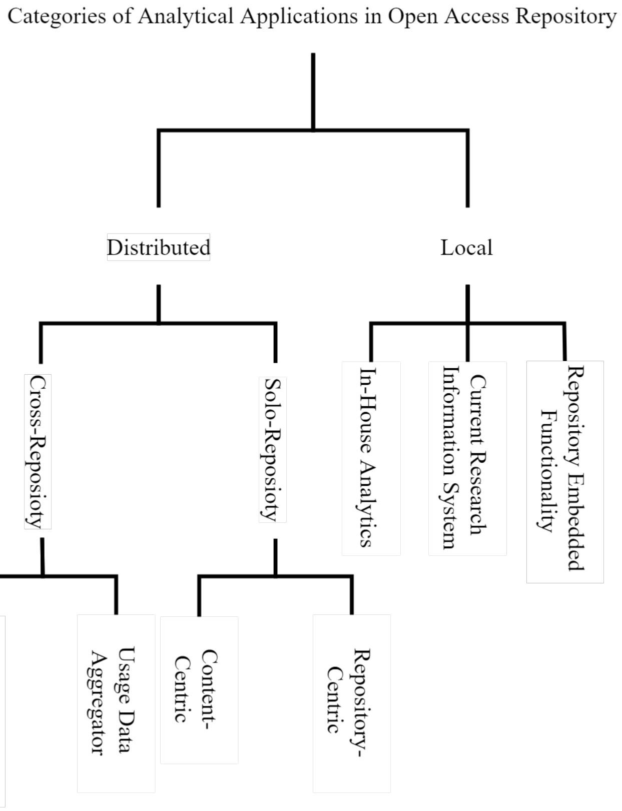


Figure 6.1: Categories of Analytical Applications in Open Access Repository

within one of these categories acted as the setting, which influenced the way the repository management interacted with them. Figure 6.1 provides an overview of these categories and their sub-categories.

The category of local analytics encompasses a set of internally operated applications, and the analytics and the data are owned and operated by and within the boundary of the institution that is operating the repository. It includes three models: the Current Research Information System (CRIS), the repository embedded functionalities and the in-house developed analytics. Yet, the distributed analytics involves entities external to the institution to operate the repository, which contributes to the analytics process. Consequently, two sub-categories have been identified: the solo-repository distributed analytics and the cross-repository distributed analytics.

The findings present three cases regarding the use of the CRIS system with open access repositories. Hence, it is part of the analytics applications operated around the open access repository. In two of these cases, CRIS was integrated with the open access repository platform. On the other hand, the third case used the same CRIS platform as the repository, as the CRIS platform was powered with a front-end layer that disseminated the research outputs as open access. In the cases with CRIS-repository integration, the CRIS ingested its content to the repository as the only source of repository content, whereas CRIS is operated in the backend.

> PURE is a dual system. It is a research information system and a repository. It could be a repository as well. So some institutions use it as differently and have a different repository in the back end but we use it as one for both. (P01)

> We have a second system called DSpace [the repository platform]... And so, everything that goes into Symplectic [the CRIS platform] is processed, and then becomes live and open via DSpace. (P06)

The CRIS can also be integrated with a system external to the institution, as it provides a wide range of analytics, both on a global scale or a nation-wide scale. For example, P(10) stated that:

> the data and metadata is created in PURE and then transferred over into Eprints, so there are analytical capabilities within PURE we can run all sorts of reports based on the metadata. Then, there is also the fact pure feeds into other Elsevier product. P(10)

What distinguishes this category of analytics applications is that it is operated by a complete system, as the repository is fed by the system instead of feeding the analytics system. In the case of CRIS-repository integration, the back-end operation of repository data is carried out in the CRIS, instead of the repository platform.

A CRIS supports a set of core purposes in the research management process. One of these core purposes is to support the decision-making process around research management (Zimmerman, 2002), unlike the open access repositories, which position them self as dissemination and discovery platforms for institutional research. De Castro (2014) draws a set of characteristics that distinguishes a CRIS from an open access repository and vice versa. According to De Castro (2014), the institutional repository and a CRIS varies in terms of the scope of data it collects. For example, while repositories collect the research output, a CRIS collects a wide variety of data related to research activities, including research grants, research projects, people, organisations, etc.

This wide range of data is a requisite of their role, which involves reporting, and analysis functionalities that support the research activities within the institution. Hence, the fact that a CRIS is specifically designed and operated to support the decision-making process and operate analytics leads researchers in the open access community to bring the open access agenda to a CRIS, and not providing analytical capabilities for open access repositories. Thus, the open access research community emphasises the role of a CRIS in terms of supporting open access publishing by powering the institution with a set of analytics tools to evaluate their open access policies (De Castro, 2019). This is aligned with increase demand and new obligations on universities to collect, analyse and report their research information. For instance, in UK universities, the Higher Education Statics Agency (HESA) and Research Excellence Framework (REF) require a detailed information on the institution research outputs and outcomes (JISC, 2016). Therefore, the adoption of CRIS is significant due to its aforementioned functionalities that can supports the institution to meet these requirements.

Also, the capability of a CRIS system to integrate multiple systems within an institution positions the open access repository as a sub system fed by the CRIS system. In the expert interviews, three cases of the repository being a sub platform fed by the CRIS system were reported and one of them utilised the same CRIS platform powered with a front-end portal acting as an open access repository. In addition to be a rich source of data regarding research activities, they were integrated with other systems within the institution such as human resources, student records and grant management (De Castro, 2019).

In addition to CRIS, the participants were engaged with a repository embedded analytical functionalities. For instance, IRStats is analytics tool quantifies the repository content, views and download. This form of analytics acts as part of the repository system components, unlike the CRIS, which is an independent system integrated within the repository. Thus, it depends on the repository platform functionalities provided by the software community.

> IRStats to the local level .... IRStats is a local plugin that sits on the server and basically does the work from the access table, stripping out all the bots and all the other bits and pieces ... IRStats, is collecting its own data as well ... processes the log files that are held on the server ... Every interaction with an ePrints record goes into an access table, which is a MySQL table that is part of the ePrints core software and IR Stats is a plugin which processes the MySQL table which is in the access MySQL table to parse it for basically stripping bots and various things out and creating meaningful statistics which it can then use to create the access and download, views and the graphs and so on. (P05)

> IRStats will justis a way to access what has gone on within EPrints itself, EPrints own records. So, it will give me Again, it tells me things like how many downloads we've had, which authors, which authors those sorts of things. P(08)

Open access repository platform developers attempt to power their software platform with analytics functionality using a set of plugins that operate for a particular purpose (ex. open access complies and download and view analytics) instead of a general analytics platform. This form of orientation can be integrated as a focus on their main role, while supporting it with a function that align with its role, such as the resource discovery evaluation function [1] and open access compliance (JISC, n.d.b). However, these embedded functionalities are limited in their ability to motivate repository users and support the decision-making process, which is a form of communication and a way for an institution to build trust with their users, as well as improve their visibility.

In addition to this form of role extension used by both the open access repository in terms of analytics functionalities and a CRIS to create an open access agenda, CRISs were interpreted as a repository with an extended data model. This, in turn, motivated Ribeiro et al. (2016) to investigate the correlation between the two. Indeed, the EUNIS and euroCRIS project examined whether a CRIS can take over the institutional repository, or whether there is an overlap between the two systems functionalities and roles Ribeiro et al. (2016). This was investigated using a survey that was administered in 20 different counties and within 84 institutions. The findings related to both project agendas revealed a negative result. They also highlighted that around 65% of the repositories were linked with a CRIS.

Beside the CRIS and the repository embedded functionalities, repository management utilises external platforms and tools that provide analytics and visualisation functionalities to exploit the open access repository data manually and on demand, based on task or through the integration of repository data with these external tools. This type of analytics practice does not use a repository platform's native functionalities, but instead operates a local service/tool fed by the repository platform. Furthermore, the role of these services/tools is to integrate the repository data with a set of visualisation and analytics functionalities. This form of services is executed by an in-house software tool or by using an on-shelf product toolset. In contrast to a CRIS, these tools are fed by the repository platform, instead of feeding the repository.

> we use QlikView [visualisation and analytics tool] to do reporting and analytics from the repository, for instance, on open access compliance around the content which we have. And that data can also be merged or mashed up with citation or other related data for other kinds of internal reporting that we can do, as well. P(05)

> We did a lot of work and decided that a third-party platform wasn't going to be suitable for the digital preservation side, for the checksums and the fixity checks, and all of those kinds of things. So, we have actually created something in-house to do that analytics work on the content. So, for the health of the content, we've developed our own in-house stuff that isn't live yet. P(07)

[1] https://wiki.eprints.org/w/IRStats2

These implementations were driven by an open access repository agenda. While the study does not cover the reason for adopting these external tools, it may reflect the limitations in the analytics functionalities and capabilities provided by open access repository software platform and the CRIS platforms. In addition, it is important to note that in this study, this form of analytics correlates with the cases where the institution operates a local department concerned with business intelligence and analytics applications. Hence, the repository data is part of the assets used to power such capabilities within the institution. In addition, it includes the on-demand analytics carried out by the repository management for a specific purpose instead of being an on-going operated service.

Although it is not only an issue of functionality and tools, as it can also be an issue of data availability. As a result of this, repository managers link their data with an external service that integrates it with external data, thus enabling new indicators such as readership, views and user behaviour. The participants reported using a set of external commercial analytics services integrated with the repository data, which enables the repository management to perform a set of analysis activities around their repository. This category includes two sub-themes of analytics: a repository-centric one, as the analytics are related to the repository; and the content-centric one, which includes analytics of the repository content, instead of the repository itself.

This type of service involves the solo use of a set of repository data without any integration with other open access repositories. However, they are integrated with other data sources external to the repository to deliver analytics. It can be classified, based on the flow of the data, into two main classes. The first of which is when the repository feeds the external analytics service to operate the external system and services, which provide analytics on top of this data. The other form of interaction involves an external data source being pulled to the repository system to operate analytics locally within the repository platform.

What distinguishes this type of analytics is that the data processing and analytics operation is external to the repository management, as well as the institution. In addition, the repository data are harvested by this external service or collected on the repository side and then fed to external system. Thus, the data sources power the analytics instead of being distributed in the processing. For instance, altimetric is one of the commercial services, provided by altmetric.com, that harvests the repository metadata and integrates it with social media, news and readership data to power the analytics oriented to the institution, in order to allow analysis of the attention given by the online community to scholarly work.

> We use the API from altmetric.com. And we use that, so that gives us some analytics in terms of attention around content that we surface in the repository through the digital object identifier. So, thinking there about

> examples within Twitter, examples of news stories, publications and policies and so on as well. So, there's that other element of analytics that we feed into our ePrints repository feeds into a wider dashboard tool for some analytics data that we use within the university. P(05)

On the other hand, the repository-centric which is not harvesting the repository content instead it recording user activities around the repository and fed it to external analytics system. This includes the web analytics services and platforms such as google analytics and Matomo.

> We use Google Analytics to understand more detailed user behaviour, why people are going through the site. So those are the three sources of analytics that we rely on. P(08)

The cross-repository analytics are mostly derived from an open access agenda on a large scale and operated and managed by nation-wide or regional entities. However, the role of the repository and its management is scaled down to data provider responsibilities such as compliance to service provider policies and standards. Thus, services use the standardisation adopted by the open access repository community to operate cross open access repository analytics. This includes the analytics that use the OAI-PMH interoperability protocol to harvest and aggregate the metadata and open access resources or the analytics that are specific to the usage statistics users' behaviour in the open access repository that uses the COUNTER code of practice.

> There's an aggregate repository called CORE, C-O-R-E. So, they gather our content from DSpace, using their OAI-PMH metadata schema. So, that's quite good. P(05)

What distinguishes this theme from the solo-repository distributed external services is that the service is specific to open access community and involves integration with other open access repositories. Mostly, the serves included in this theme has open access agenda serve the open access publishing as overall instead of a particular repository *per se*. For instance, the IRUS UK a nationwide analytics service aggregates the UK institutional repository usage statistics to reflect the growth of green open access in a nation-wide level and its uses.

> The closest nationwide service that we would plug into is IRUS UK, so the service that downloads the ... well, aggregates, downloads statistics of multiple repositories. So we've got Pure talking to the IRUS UK servers, and I think every morning at nine o'clock we send up to their service the latest stats. P(02)

6.3.2 Section (2): The Participatory Roles of Repository Management within Analytics Applications

The participants reported a set of interactions within the analytics applications highlighted earlier. These interactions are categorised into a set of roles based on the nature of interaction made. The final set consists of four roles: the analyst role, the administrative role, the data and system management role and the system development and support role. The full list of themes and sub-themes that assisted with their illustrative quotes are provided in Appendix E.

The repository management engages directly with the data or with the analytics system/service/tool as data analysts, in order to generate insight and support their role at the institution, and in the open access repository or open access publishing in general. Part of these analytics are concerned with the adherence of open access repository content to fund open access policy such as the Research Excellence Framework (REF) open access policy. Therefore, the repository management practice the role of analysts by evaluating their repository content as users of analytics services.

> The availability of open access items, for example, within the repository. Whether they adhere to the REF open access policy, which is a big thing at the moment. P(01)

On the other hand, they engage in the complex task of analytics that involves data gathering from external sources, integrating it with repository data and processing it to find insight that support the decision. An example of this process is provided by P(10) who described the process of examining the adherence of National Institute for Health Research (NIHR) output to their policy.

> Sometimes ... the data that [is] in pure, and therefore that in prints, is sometimes incomplete because it rely on our authors to upload it. So, some of the places to look on is Scopus, a big abstracting and indexing service... what we do [is] run jobs where we pull new content that has our [the institution] authors on it from Scopus into pure, which then populates it to the eprints ... in terms of analytics [...] we sometimes looks at Scopus ... download, usually as CSV, a list of papers. Say for instance, if we're doing a return to [National Institute for Health Research] NIHR where we're saying we want to look at the last three years worth of content published as a result of NIHR work and whether that is open access. So, what [we] might do, we get a list of authors associated with NIHR, write a billion logic search in Scopus to find all their content or look only at the [the health related research centre in the institution] or something like that. And then we pull that down, but the

> thing is, we are then left with information such as these DOI title authors. So, it is not really very helpful, Scoups, in term of looking on open access. So, what we then do is maybe run a search in a service called Unpaywall, which integrates some of this information drawn from eprints ... but basically, you can use that to tell you some interesting information to inform decisions that you make ... but if you really want good quality information about what? Gold open access? Green open access? where people are publishing? It is easier to get it from unpaywall by doing a search on the DOI, so unpaywall search all the institutional repository and uses Crossref and other services to pull together massive data and information ... So, if you send them a list of DOI that you want to look at, they will send you back a list of DOI with information on whether it is open access and what sort of open access it is. So what you can then do in excel is apply filters or use a private table to basically say 'Is the journal fully open access? Yes or no?'. And then you can say 'Is the article open access? And if it is. P(10)

Beside the analysis of the adherence of open access content to open access policies, the repository managers analyse repository users' behaviour and the online communities' attention to a particular scholarly output. In addition, the repository management interact with a set of tools to gather insight into repository users' behaviours including their download or viewing actions. This is facilitated by repository platform functionalities or well-known commercial analytics product such as google analytics, which may include both human interaction or bot machines. This process can be very specific to the repository or utilise external services analysis indicate which particular scholarly documents have received the attention of the online community. In turn, this provides alternative indicators of bibliometric indicators that are used in the citation count, which evaluates the performance of a particular scholarly output, author or academic department.

> It [analytics service] provides a mix of downloads and also access. So, both of those things aren't the same. So, it's really interesting to see where the interest is versus where the downloads are. P(05)

Benchmarking the repository and institution with other repositories and institution is another form of analytics task carried out by a repository management, as they engage with nationwide, regional or international analytics services to position their repository in the open access repository global map. This also can take place at a content level, where they drill down to uncover the content distribution over the open access repository infrastructure and how it is being used across repositories.

> You can go onto the IRIS-UK website and you can compare yourself against other organisations. P(02)

> It [the analytics platform] will allow you to export it as a CSV and manipulate it as you like, to produce statistics and for other insights, which the top universities like being able to benchmark themselves against each other. P(07)

> It's starting to look at ORCID which are embedded in repositories and so on as well, so there's a very interesting analytics piece around the data, which IRUS-UK is capturing at a national level, [and] you can then look at it and go, 'ah right, yes, this DOI appears in however many different repositories'. Or someone with this ORCID has content across a multiple number of repositories and what do the downloads and the related statistics that are aggregated around look like? P(05)

In a broad sense, repository managers use analytics and play the role of analysts to support the decision-making process during repository management, however, they are not limited to this case. They communicate this insight gathered from their data internally with the academic department, institutional department or externally with the funding bodies, national pressure groups or individual academics to support self-archiving practices.

> We can also run internal reports to feedback to colleagues about various things, such as issues around items that may not be compliant for the next ref exercise. So, we can run internal reports like that as well. P(05)

> ... to inform funding bodies and national pressure group. P(10)

> And that's the group that we provide reports to every two months, to tell them how compliance is going with the ref policy here at the university and if there's any issues with that. But we use QlikView for that, because it's got the nice visuals. P(05)

In the case of the analytics carried out through a CRIS, the open access repository acquires its data from the CRIS. Accordingly, the repository content and metadata administrative tasks take place in the CRIS system, instead of on the repository platform. Thus, the administrative role of open access repository management is transferred to the CRIS. This leads to a set of administrative interactions being made by the repository management within the analytics system (the CRIS). This includes the metadata quality control and content management.

> Basically, anything that goes live into our system has to be checked by a human. P(06)

> I work with the day to day operations team, so we are responsible for running the [unclear], so we get the best metadata, full text up that we can, which of course informs the various [unclear] from that. P(04)

> There are a few minor side-effects, because it automatically goes into the repository and we could probably technically fix it, but sometimes we end up with duplicates that need to be de-duplicated. P(03)

A part of taking the administrative role in the locally operated analytics system, some service provider that operate analytics as added value services power the repository management with a set of functionalities to manage their harvested content remotely.

> We can go into Core if we want, perhaps to take a document down that's been harvested by them. So, we have that kind of interaction. P(05)

The repository management take the role of coordinating the analytics service users, in order to encourage them to supply the data for external third-party entities. Thus, it is an intermediary between the data owner and the analytics service provider.

> ... facilitate or coordinate the user groups in the university within the three colleges. P(01)

> What we would do then is to encourage, because researchers have to upload or to sign off on the submissions themselves on the Researchfish. P(01)

The position of CRIS to open access repository transfers the administrative responsibilities from the repository platform to CRIS platform. Hence, the administrative interactions of repository management team with the CRIS systems is identified as one theme of interactions between the repository managements and analytical system that is exploiting the repository data. This includes the management of policy around CRIS, as well as the management of its processes.

> I would say we manage the policy side of it. We manage the processing. We respond to authors and users, external users. P(06)

The repository management takes the role of data manager, as they are the owners of the open access repository data used in the analytics process. Thus, they take the responsibility of liaising internally with other departments within the institution or externally with other services.

> [The] team helps people to get data off of [the repository], maybe individuals want to know 'Can I have a list of my publications?' or they want to have a list of someone else's publications, then the team will help people do that. P(03)

> We are definitely providing them with statistics, as are dozens and dozens and dozens of other universities. P(02)

> We also produce the data for, we have what we call a research, planning and strategy committee here. P(03)

> ... external liaising the data with external groups. P(10)

Also, the findings reflect that a set of technical and development support were made by the repository management towered analytics services. That included adding new functionalities or providing analytical information on demand.

> If there's any requirements or new add-ons to the service, whether they are technical or otherwise, then we'll decided whether or not those need to go in. And usually we'll resource putting them in and setting them up and making sure it's integrated with the technical side and also the process side of things. P(03)

> Or also, dealing with any queries, so for instance, we have had queries if statistics aren't available, what statistics we can do, some local work to create some analytics that we can provide on an individual basis. P(05)

Also, they were found to be aligned with their commitments to comply with the set of standards used by the analytics service provider. This was shown to include a close collaboration of the repository technician with the analytics service provider. It was also found to include the monitoring of the system status and the troubleshooting of analytics system technical issues.

> So there was a bit of interaction there, but that was just a case of us trying to fix it at our end. They weren't necessarily coming to us and talking about the sorts of reports that they could do. P(02)

> So the role I had there was to work with the IRIS-UK to work out why our stuff wasn't making it to their servers. So there was trouble-shooting at my end and trouble-shooting at their end. P(02)

> It came through as a statement from the EU saying that you should be OpenAIRE compliant if you receive funding from them. And we checked our repository and we felt it was, but that we would just, we just mapped the fields we had and then went back to them and they checked and said, 'yes, you are compliant'. So, yes, its just there now, we don't really have any ongoing interaction, other than hearing occasional news items about developments and that aspect. So, it works, it's there. P(03)

In addition, they were actively engaging in the development stages of analytics services development. For example, P(03) reported a pilot analytics project they were involved with that launched analytics services on open access repository data. Also, they provided feedback to the service providers based on their position as analytics users.

> JISC were doing a pilot of this and they asked us if we would like to be involved. So, we were involved quite early on, maybe even the first site that did this. And they just said, 'Are you interested in doing this?', [and] we said 'Yes'. P(03)

> So they've done a bit of work on how the statistics look. And they have sought feedback. I haven't fed back to them, but they have at least been proactive and come to us and said, 'Right, so we can now do this. What do you guys think? P(02)

Regarding open access analytics, which is usually carried out through cross-repository analytics, the role of the repository and its management is scaled down to data provider responsibilities such as compliance to service provider policies and standards. Thus, their interactions are scaled down and limited to minimal interactions, however, this interaction is constrained by a set of coordination techniques, methods and strategies.

The participants highlighted a set of these techniques, methods and strategies that are used by the open access repository community to coordinate with external analytics. They include the following:

- OAI-PMH Protocol: The Open Access Initiative - Protocol for Meta-data Harvesting (OAI-PMH) which is a low barrier **interoperability framework** based on meta-data harvesting (discussed in the ROAR case study in Chapter 4).
- RIOXX: RIOXX is a **metadata application profile** and set of guidelines developed to support the Research Council UK (RCUK) to monitor compliance to open access policies. The RIOXX is composed of a set of guidelines and metadata format that enables a standardised exchange of metadata, as well as normalises

repository managers interpretation of common metadata elements. RIOXX is used by nation-wide aggregators such as CORE, as well as regional EU OpenAEIR aggregators.

- COUNTER code of practice: A **code of practice** designed to support the recording and exchange of usage data statistics, with the aim of providing a consistent, credible and compatible approach to reporting usage data.
- DOI: The Digital Object Identifier (DOI) system provides a **persistent identifier** that uniquely identifies the digital object, which is standardised by the International Organisation for Standardisation (ISO).
- ORCID: The Open Researcher and Contributor ID (ORCID) is a **persistent identifier** that uniquely identify the authors on a worldwide scale. Thus it is used to identify the authors across repositories.
- Resource fingerprinting engine: a software solution that identifies full-text research through the generation of **resource fingerprinting** using text-mining techniques.
- Plan S: an **initiative** for open-access science publishing that aims to support open science in general. Its lunched by cOAlition S and is supported by the European Commission and the European Research Council (ERC). As part of these initiative, a set of mandatory criteria and recommendations for open access repository were developed, with the aim of supporting the standardisation of the dissemination process of open access research and repository metadata.

The open access community incorporates these techniques to support each other and foster the open access vision. While the persistent identifier enables the services to identify and coordinate the analysis of the data cross-repositories, the interoperability protocol, metadata application profile and code of practices coordinate the data exchange between the open access repositories and external service providers. Furthermore, repository managers are motivated by open access policies that mandate the adoption of coordination techniques and strategies.

In addition to these coordination techniques, the open access repository community uses a set of communication channels to communicate updates, changes and service requirements. These techniques include: emails, mail lists, helpdesk systems, mediated communication, platforms and conferences. The participants highlighted direct communication via email as a common communication method used when approaching a service provider or vice versa.

> I've also been dealing with IRUS-UK, so I know the people there. I may also be the contact for OpenDOAR, ... we have a number of named contacts. A number of named contacts and a number of routes of access in order to

> escalate and to deal with, whether we're dealing with a distributed vendor or we're dealing with local issues. P(05)

Although in some cases, communication is not direct with the service provider, but is instead mediated by a nation-wide representative. For example, the OpenAIRE service provider has national desks across the EU, and JISC is the representative for this in the UK. Thus, any form of communication is mediated by this national desk and the repositories, instead of there being direct communication with OpenAIRE as the service provider.

> OpenAIRE has a national desk, and the one for the UK is with JISC, actually, [...] he [a European open science manager] contacts us to say, we are having to make sure that OpenAIRE is getting the stuff that the EU funds from your repositories into OpenAIRE. Can you test it?. ... If we have problems with OpenAIRE, then we can contact JISC, who is the national representative. P(06)

Another form of communication that commonly occurs is via a mailing list, whether from a nation-wide group or a service specific group. In nationwide groups such as UKCoRR and OA Good Practice mailing list, the service provider is on the mailing list, and updates and the changes are communicated through the mailing list.

> I suppose there is a mailing list called OA Good Practice, open access good practice. That's a JISC mailing list. And that allows people to discuss the requirement policy, new requirements, which there are many. There's also a few things coming through called Plan S and new funding requirements. So, we can discuss on there what people are doing. P(06)

> So sometimes I will get a direct team email, because they have a record of who to correspond with, that allows the most reliable method, because the staff turn over ... so there is also a community mailing list, ... we subscribe to JISC mailing list .. and everyone equivalent to myself [repository manager] or equivalent to [repository administrator] will probably be a member of that JISC mailing list ... and there are people [who] represent these organisation [server provider organisations], they will be in this mailing list, so usually if there is big things to change, they will send an email [to notify about the change]. P(10)

> We have a Symplectic user group, which is a European user group, that's worldwide. P(06)

Also, the community utilises the conferences as an approach to communicate their changes and developments.

> There's actually a conference coming up on 5th April ... 5th July, at the University of Liverpool. So, we have a user forum. We have a mail list ... And we ask for changes. So, we have feature requests, if we want something to be changed, or if we have problems with dates, or… Because we have all these REF 2020 rules, we also need reporting mechanisms within Symplectic for REF 2020. So, we either use the web forum to request things or feature requests, and also ... we communicate with each other. P(06)

6.3.3 Section (3): Concerns Related to Analytics Applications on Repository Data

The participants brought to the researcher's attention a set of concerns around the analytics application of the repository data, which can be grouped into three main categories: **data-related**, **analytical** and **cost concerns**.

With regards to the data-related concerns, the participants stated that they had no concerns regarding the exploitation of repository data for analytical practices in a broad sense, as they believe that open access means being accessible without there being any barriers to the data. In addition, they argued that it is part of the university's responsibility to make repository data accessible, because it is a public data. With regards to the unlawful use of data, they stated that they place confidence in the policies they adopt to protect their data, which takes the form of institutional level policy or nation-wide law enforcement. Also, they expected the user of the data to use the data based on the license adopted for each repository item. Below are some examples of participants statements on the subject:

> Because it is an open access portal that we are providing, most of this information is available to the public anyway. It is our responsibility as a university to make this information open to the public. P(01)

> If there is a CC-BY licence, you can do that, but we do have also the University of [the institution name] accepted manuscript licence, which is a very liberal licence, but also may take the rights for the publisher. So if the publisher does not specifically say you may make this article open access green open access using CC-By license [..] if you don't specifically say that, then we will apply the University of [the institution name]'s accepted manuscript licence, which says you can access this, you can use it, but you can't host it on another website. And [there] may be… probably [are] copyright licence

> fees to pay for the reuse of materials, and you need to check the original licence. So if the original licence is copyright, Wiley [the publisher], so you expected to check what Wiley licence allow you to do, so in which case, it may restrict you in what you able you do, in terms of text mining. P(10)

> One of the things that the original guidance for the plans said was that everything in the repository should have machine-readable licences. P(10)

In addition, participants reflected concerns regarding the commercial use of the data, even if it is for analytics practices.

> And also, you're a commercial company. You're making money from us. Yes, it's a service that exposes our research and our theses, which is good, but you're also making money from it, and that's not our business case. Our business case is to make this stuff available for free. P(06)

> For research. Not for your stockholders to make money. So, we had to say no. So, that was the worry that we had, was that they wanted to harvest our stuff, and they were making money from it. P(06)

Furthermore, the participants distinguish between the form of data used for analytics as the repository may have sensitive data and other public data. So, the level of concern varies, depending on the sensitivity of the data. Sensitive data is brought to repository data through analytics practices, where the repository data is integrated to provide insights into REF compliance, which is personalised based on user data.

> But some of the data could potentially be a bit sensitive, for example where someone has explained to us that they were off on a long-term sickness and that is why they could not comply. Then you could be getting into personal data concerns and so that's why we've taken the step to lock that down a bit. P(03)

In this case, the repository management is a controller and the collection, integration and analysis of the data are regulated processing under GDPR (see Section 5.4.1.4) and should be carried out under lawful-bases determined by GDPR legislations. Hence, the repository management has taken further actions to comply with such regulations. This includes transparency with the data subject regarding how the data will be processed and power them with capabilities to exercise the right to be forgotten. Although the statistical purpose is exempted from the purpose-limit principle on the GDPR, the proper protection and safeguarded is critical.

> In our internal QlikView application for REF compliance, we have restricted it to limited users and you have to attend a training course to explain to them to be careful, and there are certain parts of the data that are not for public consumption. P(03)

With regards to the cost concerns, the participants highlighted the commercial services that use their data for analytics practices and engage in the process of integration and compliance to service provider requirements. Accordingly, the service is provided as a commercial paid service. P(10) provided a description of this situation with the situation *"giveaway"* literature and paywall, which is enforced by the commercial publisher. The commercial service providers integrate their commercial services with the repository data, which is curated, integrated and managed by the repository management team, where the service provider adds value to the data at a cost. However, the main concern is not due to the added cost, but instead of the relevance between the cost and efforts placed on the benefit and value they provide, and whether it is worthwhile or not.

> Quality, availability and also return on the effort... return on the investment. When I say return on effort, or effort, I don't mean in the same sense, I mean when we pay for research. I mean in terms of was it worth spending 20 hours of one of my staff member's time making sure that it was up-to-date. We can't see the benefit of it and some of it you don't even know until you have done the work. P(10)

> It is not a concern so much, it is an acknowledged madness, so we pay Elsevier to provide us with the repository, we pay Elsevier to provide us with Scopus, we pay Elsevier to provide us with SciVal, we update information to goes into Pure and we go through cleaning processes where we clean this data in SciVal and Scopus to make sure our Scival analytics are as accurate as possible, ... we have to go through and clean up the data to make sure when we have reporting at local, national and international level that the information is accurate and as clean as possible ... So we are actually cleaning up the data for Elsevier, who then sell it back to us [...] We are doing a lot of work, they are doing a lot of work. They are definitely adding value. They are definitely presenting it in usable fashion, but is the value of what they provide so consistent with what they are doing. P(10)

Furthermore, P(02), P(03), P(06) and P(10) stated that this form of service results in new responsibilities, complexities and effort needed, in addition to the primary responsibilities they have as repository managers. This is added to by the existence of a variety of systems, services and tools that have developed that have a range of dependencies and requirements. As a result of multiple service providers and entities enabling the analytics service, repository managers have to manage a number of relationships, as well as coordinate and communicate with these service providers.

> This is the negative side, which is there are many relationships to manage, there are many different system written in many different coding languages ... So, there is big management thing there. You have to keep up on top of the information. You have to make sure you are receiving those emails. You have to make sure you are reading hundreds of those emails that are coming from a subscribed email list. You need to win time and work with the IT solution department to make the new changes. It is difficult for some services to say that it actually worth your time and effort ..., especially when you are not using this information on a regular basis. P(10)

With regards to analytical concerns, the analytical insight generated by the analytics systems and services influence the decision-making process for the funders, the academic institution and within the repository management. Thus, the accuracy of this insight and representation of the insight is one of the concerns highlighted by the study participants. Moreover, analytics is used to benchmark repositories and institutions nationally, regionally and/or globally. Thus, the consistency of benchmarking criteria and the transparency of the analytics process also concerns repository managers, as crucial decisions are made based on these benchmarks.

> As we discussed a lot about how clean the records are, how accurate the records are, how consistent the records are between different institutions ... and then the fact that the information is being gathered from Pure to go to eprint ... I am a little concerned that this information is not complete, and it not accurate, and we are using it to make a big decision. There are three things; the local level, nation-wide level and international level. At the local level, is the time and effort worthwhile? And is there a return? Are we going to get more funding as a result of this? Are we demonstrating how internationally important we are as research institution? Therefore, are we going to get international collaboration as a result of the work that we are doing? So we have to prove that. On national level, you have funding councils they want to understand the return of their investment, how impactful the research has being and you have multiple measures of impact that you have to report on, you have that responsibility ... We have to understand the limitations of the data, we have to communicate the limitations of the data, ... And then on the international level, you again want to prove the return of investment ... quality, availability and return on effort. P(10)

In the light of these concerns, repository managers' reactions to these concerns were sought. The participants highlighted that an authentication system and access control is adopted to control who and what type of data should be exposed and used. These authentication systems function as a reaction to the data concerns, including the integration between the two forms of data within their service, which is associated with locally operated analytics. Also, system and network monitoring tools are adopted to monitor the harvesting and automated full-text download of their collection and restrict it to those who have permission. With regards to the analytics conference, the signing of DORA declaration, the development of the comprehensive research responsibility matrix, and guidelines to the users of the data on the limitation of the data have also been adopted to ensure the accuracy of the data analytics that uses repository data.

> The data that's on QlikView for REF compliance, we have locked that down and made people go through training and it's only limited registered users. Mainly that's a precaution, because the vast majority of the information, there's no concern about it. P(03)

> We're developing a really comprehensive responsible research metrics policy, so there is... the international reaction has been DORA in San Francisco, and the Leiden manifesto, those ask people to make sure they are using accurate data and the thing is reproducible and using appropriate metrics. We decided to write our own policy and also to sign DORA. The idea of having our own policy is that we will be entirely explicit on what is and what is not acceptable ... you have to be assured of the quality of the data you are using, and provide provenance on the data and, well, description ... At least the people understand the limitation of the data. P(10)

> We are developing guidance alongside human resources and faculties to make sure that the people understand what is and what isn't appropriate for the research analytics. P(10)

6.3.4 Re-visiting the OAA-OARD-SM Framework

The findings of this study necessitate the re-visiting of the OAA-OARD-SM conceptual framework to reflect these findings. There are four considerations highlighted by this study that require the attention of the conceptual framework. Firstly, the fact that the study reflects a variety of types of interactions by the repository management regarding the analytical applications of the repository data provides an insight into the awareness, to some extent, of the complexity of analytics, and the availability of the skills required to deal with analytics practices by the repository management. The skills and the complexity of the task is one of factors that influences the operation of social machine processes (Satzger et al., 2013; Yang et al., 2016). While this study does not aims to examine these skills, the insight acquired calls for such analysis. Secondly, the participants' perspectives of the correlation between analytics interactions they made when using analytics applications and their role as repository management is emphasised positively, as they argued that it is one of their responsibilities. However, there is tension between the need for the evaluation of the revenue of investment based on the efforts they make to support analytical practices and their commitment, due to the fact it is one of their responsibilities.

> I don't think they add to my responsibilities, I think they are part of our responsibilities already. I think it is expected, because we understand parts of the application and how the application works and the data models that the application uses. We would be the best people to know how to get that data out for our users. P(01)

> It's important that this is something we know about, and it's important it's something that we check periodically, to make sure it's working. But other than that, I wouldn't say it's substantially adding duties and responsibilities to my current job. P(02)

According to Smart (2018b), the 'perennial problem' of online communities in general, including the social machine, is how to motivate people to take part in the system process. In the case of open access analytics and repository management, this study provides insight into the established responsibilities and motivations to take part in it, whether it is at a local, national or international level. Hence, it can be argued that what is challenging is how this effort and awareness can be harnessed to establish fruitful interactions within the community. Shadbolt et al. (2019) stated that;

> When a social machine lacks a central coordination body, then its survival and reproduction through time depends on it being the focus of a fruitful interaction with its community. (Shadbolt et al., 2019, p 44)

To gain further insight into this, the concerns around analytics applications using open access repositories can be incorporated with the opportunities offered by OAA-OARD-SM. The multi-systems, entities and requirements imposed by the service providers and open access community to operate analytics functions add an extra load to the repository management role. Consequently, participants called for an evaluation of the worthiness and revenue of such interactions. This concern has emerged due to the various dependencies, standards and requirements of the service providers.

On the other hand, social machines can be overlapped, that is to say, a general social machine can be used to create another social machines process (Smart et al., 2014). In that way, ad-hoc communities can be created within overall larger social machines. Furthermore, Murray-Rust and Robertson (2014) proposed a framework that enables a social machine process using an existing infrastructure to reduce the cost of coordination. In addition, it also decreases the expense of joining a particular task process through a shadow institution and shadow agents [2] to guide the participants in the coordination of a particular real life task, whilst also taking advantage of the existing community and infrastructure.

They demonstrate their framework on twitter, where the twitter hashtag is used to create ad-hoc communities. Given that the OAA-OARD-SM adopts the web observatory as an infrastructure to support analytics functions and harmonise the technical complexities, service providers can be engaged in analytics by forming an ad-hoc community around their service. Technically, web observatory implementation incorporates a “project” as a grouping technique that enables a number of analytics and datasets to be catalogued within one project (DiFranzo et al., 2014). Thus, service providers can mirror and be represented by a shadow institute that guides the repositories to join and collaborate in the realisation of their analytics.

Thirdly, Murray-Rust et al. (2015b) use the concept of wayfaring[3] to interpret the phenomenon of social machines, explaining that a social machine provides a landscape for people to navigate. Where this landscape is created by the participants and enables the participation and exploitation of another participant. Furthermore, social machines can be designed to identify tasks carried out by people and identify incentives that support their engagement, as well as identify the circumstances that support inhabitation. They also ensure this landscape should not just support transport from A to B, but instead the wayfaring experience should add new meaning to the landscape. According to Murray-Rust et al. (2015b):

> In wayfaring view, on the other hand, a journeyer is situated in a landscape, with signs which can be read, and possible directions to explore. Rather than a top-down map of the world, on which routes can be meticulously planned out, navigation is local and responsive. (Murray-Rust et al., 2015b, p 1144)

Within the OAA-OARD-SM, the analytics task is conceptualised to be carried-out within the individual repository taking into account the analytics goal and analytics specification, including the data quality specification and goal specifications, where, the

[2] a shadow agents hereby are a digital mirroring of real-word participants to guide and coordinate the process.

[3] Murray-Rust et al. (2015b) uses this term with distinguishes between the transport that characterised by the purposeful structured movements and the wayfaring which is about navigating through landscape reading signs and making paths for others.

web observatory is the landscape to create analytics task and support it with the analytical application. Taking advantage of the aforementioned the analytics exploitation's classified under the local theme, the OAA-OARD-SM adopts the wayfaring view to externalise the data analytics process to the web-based landscape, which enables the analytics assets for open access analytics to be co-created and the output of analytics practices can be reused and replicated. This enable the participants to react locally instead of following an overall top-down plan to accomplish their own open access analytics agenda (for instance, to understand their own repository compliance to open access polices). Hence, their actions are contributing to understanding the open access analytics agenda in greater scale.

Beside their actions contribute to the development of the landscape (the web observatory). To clarifies, the open access repository management teams is situated on web observatory as landscape adopting the goal specifications and data quality specifications, linking their repository data with web observatory platform, using and reusing its analytical apps to realise a particular analytics task. This process enables the participants to drive new specifications for another analytics task using similar analytical apps or derive new specifications for new task realised with different apps. In Figure 5.9, the analyst interacts with analytical apps through a set of explorations process enabling new specifications used to interact with analytical apps for another explorations loop. This loops of explorations allow the participants to identifies new task.

In this scenario, the participants of the system are not users instead they are participants engaged with the system processes and contributes to its developments as whole. In addition, the overall setting enable the participants to stimulate a wayfaring approach instead of fixed specifications realised by the system and its participants. Beside, their participation is not guided by plan instead of their localised actions motivated by their need to understand their repository data.

Fourthly, this study identifies a set of analytics purposes drive the repository management to use and interact with the analytical service. Figure 6.2 highlights these purposes. In addition to the operationalisation of the wayfaring concept on the social machine phenomenon, Murray-Rust et al. (2015b) emphasise on the concept of 'entangling' in social machines, showing that when designing the social machine, the participant paths need to be designed in 'entangling'. In social machines the participants' paths are their activities in the social machine as a landscape. Accordingly, in this study the participant path is operationalised by the OAA-OARD-SM conceptual framework participant's activities to achieve analytics purposes.

These purposes are naturally entangled for example, repository benchmarking and cross-repository content distributions monitoring. Furthermore, open access analytics is an entangled process, as the open access deposit activity monitoring service, is entangled with the activities of monitoring an individual repository to gain insight into open access

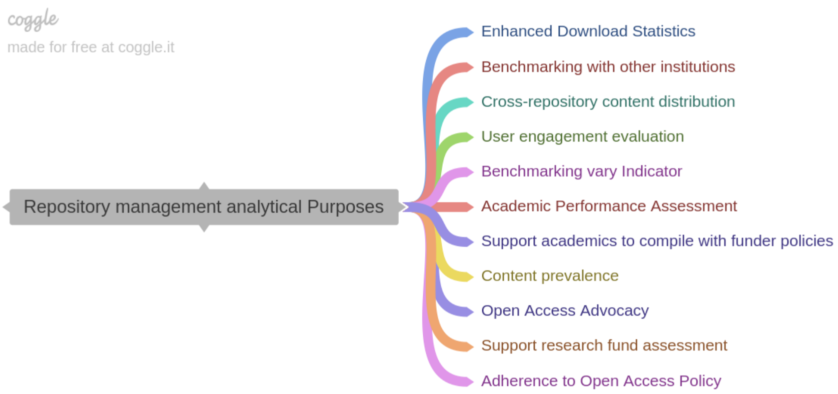

coggle
made for free at coggle.it
Repository management analytical Purposes
Enhanced Download Statistics
Benchmarking with other institutions
Cross-repository content distribution
User engagement evaluation
Benchmarking vary Indicator
Academic Performance Assessment
Support academics to compile with funder policies
Content prevalence
Open Access Advocacy
Support research fund assessment
Adherence to Open Access Policy

FIGURE 6.2: The repositories management analytical purposes identified during the expert interviews.

repository deposit dynamics (Carr and Brody, 2007). In addition, in a more complex process, the estimation of the effectiveness of open access policies is based on the effectiveness of the entangling of the open access mandating conditions within an individual repository, as the degree of effectiveness is examined across repositories to score each open access mandating condition effectiveness (Vincent-Lamarre et al., 2016). Based on the second and fourth considerations, the data analytics layer is modified. The OAA-OARD-SM conceptual framework positions the analytics goal as the central concept to the data analytics layer, without any explicit proposition between the analytics goal and how this goal correlates with the open access repositories management. While, the central position of the analytical goal in relation to the layer is still intact, the open access repository management purposes and services provider purposes need to be central in the designing of the analytics tasks (see Figure 6.3).

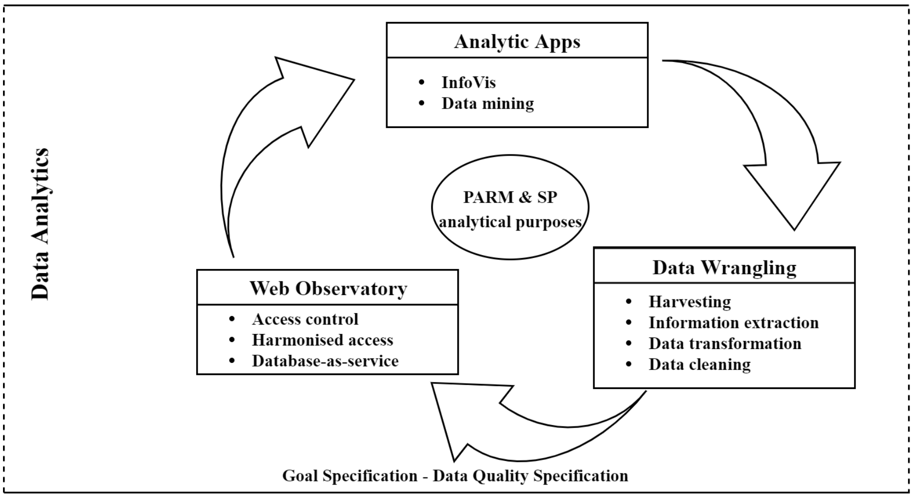


FIGURE 6.3: The OAA-OARD-SM data analytics layer after re-visiting.

6.4 Chapter Summary

In this chapter, the exploitation of open access repository data for analytics parasites at the repository level was investigated. Based on this investigation, five main themes were identified, as well as four interactions themes made by the repository management within the five exploitation themes. Also, several concerns highlighted by the participants during the expert interviews were classified into: data-related, analytics-related and cost related concerns. These concerns, along with the model and interaction, were utilised to revisit the OAA-OARD-SM conceptual framework, which supports its composition and call attention to understanding the repository management analytics purpose as central to the social machine process which can support the open access analytics.

Chapter 7

Conclusions

7.1 Introduction

Open access publishing, and open science in general, have undergone a number of developments and key milestones that have shaped its principles and scale its community (Suber, 2009). They have also taken advantage of a number of opportunities offered by promoting openness, as well as considered gaps in scholarly communication encountered by open science technologies (Lynch, 2003; Brody, 2006; Harnad, 2008a). Along with advocacy activities, which have also played a major role in the adoption and development of open access publishing. What distinguishes this advocacy is its research community, which supports it with evidence of the open access publishing advantage, adoption and coverage (Suber, 2006; Brody, 2006; Swan, 2010; Pinfield et al., 2014; Vincent-Lamarre et al., 2016). A number of these efforts are data-centric efforts and in some cases provided as an ongoing service.

This thesis takes it a further step by consolidating these practices into the notion of 'open access analytics' to bring the issue into focus and support the communication of the idea. Hence, the data analytics literature has been deliberately reviewed with the aim of connecting it with the open access agenda. This was achieved through a narrative literature review on the open access publishing concept and data analytics concepts, which led to a synthesis of the concept of open access analytics. In addition, open access analytics were defined using Cooper (2012b) framework of the characteristics of data analytics. Therefore, in a broader sense, open access analytics refers to a data analytics process characterised by open access publishing as analytics object despite to the analytics subject is used in the analytics process. Yet, with this definition, analytics of the open access agenda is seen as broad topic with challenges to investigate.

Inspired by ROAR analytics, the thesis scaled down its scope to the analytics of open access repository data, specifically by investigating the delivery approach of open access analytics on top of open access repository data. Accordingly, the review was extended to cover open access repositories and their data, demonstrating their position in open access publishing and how the open access repository infrastructure supports the open access analytics functionalities. The review emphasises the strong links between open access publishing and open access repositories as a strategic route to make research open access, its ability to nurture a wide range of open access research and the wide adoption of its infrastructure. In addition, it is supported by a set of approaches that enable the development of value-added services, including analytics services. These approaches depend on the interoperability layer adopted by the community, which allows an automated data exchange between the repositories. This is where the widely adopted approach is the OAI-PMH interoperability approach, which uses the data-service provider approach to operate the value-added services.

Being the dominant approach used to operate value added services on top of open access repositories, including open access analytics, this doctoral thesis contributed an explanatory case study of ROAR analytics to demonstrate the fact an OAI service provider's conventional approach supports open access analytics. The exploratory case study explains how open access analytics requirements can be realised using an OAI service provider, including the defragmentation of open access repository infrastructure, the operationalisation of the unit of analysis, data collection, data reduction and normalisation and the information visualisation. Furthermore, the case study demonstrated the limitations and challenges associated with OAI service provider's approach, including a single point of failure and sustainability issue of the service provider, quality of OAI-PMH adoption, processing and network bottlenecks, huge volumes of data and long-term management.

This doctoral thesis incorporated a ROAR case study to motivate further investigation promising yet not investigated concept in the open access repository context, which is adopted from the web science community. The web science research community proposed the concept of social machines to interpret and understand the new form of large-scale collaboration mediated by web-based technologies. In addition, the concept has been proposed to the confront challenges faced by other research communities (Byrne Evans et al., 2013; Murray-Rust et al., 2014; Tiropanis et al., 2014b; Ahlers et al., 2016). Accordingly, this thesis provided a conceptual study framework combining the concept of social machine concept and the open access analytics process into a single conceptual framework, denoted as OAA-OARD-SM. The open access analytics process is identified using conceptual analysis, and synthesis using a social machine view. The OAA-OARD-SM conceptually demonstrated that open access analytics is joint collective collaboration mediated by the web observatory infrastructure, which is driven by open access repository management interactions. In addition, it illustrates the fact that the open access analytics process occurs across four layers, namely the open access repository layer, open access registry layer, data analytics layer and open access analytics layer.

One of the main propositions of the OAA-OARD-SM conceptual framework is the role repository management has in the realisation of open access analytics, with an emphasis on the opportunities of their position in the open access repository infrastructure. Accordingly, this research confronted the limitations of previous research on data analytics practices within a particular repository boundary, by conducting a qualitative study using expert interviews to explore the exploitation of open access repository data for analytics practices by the repository management team, who are regarded as experts. The study gives further insight into the type of analytics applications operating on top of open access repositories, the interactions made by the repository management regarding these applications, coordination practices with external service providers and the concerns highlighted by the study participants around these applications. These insights were incorporated to support and modify the OAA-OARD-SM conceptual framework.

In overall, this research attempted to establish a connection between the open access repository, the data analytics and social machine concepts. It demonstrates the analytical value of the open access repository as data sources to carry out the data analytics. Hence, the data analytics is expressed as the process of utilising the data sources in order to obtain actionable insight through problem definition and the application of statistical models and analysis against existing and/ or simulated future data. The correlation between the open access repository as data source is established based on its linkage with analytical object (open access publishing) and the availability and accessibility of rich data about analytical subject. Due to the domain-oriented nature of data analytics concept, this connection is consolidated into the notion of “open access analytics”. Accordingly, this doctorial establish another linkage between the concept of open access analytics and social machine taking the advantage well-form open access repository community represented in the repository management teams to re-frame the analytics into social machinery process.

7.2 Thesis Contributions

First, this thesis contributed to the literature that links the open access repositories with analytics practices. The open access repositories as open infrastructure is investigated for varying forms of information services highlighting the infrastructure and system (Knoth and Zdrahal, 2012; Liu et al., 2002; Robinson and Horstmann, 2007; Müller et al., 2009), analytical approaches and tools (Shadbolt et al., 2006; Carr and Brody, 2007; OBrien et al., 2017; Vincent-Lamarre et al., 2016), and protocols and standards enables these functionalities (Bell and Lewis, 2006; Reese, 2009; Lagoze and Van de Sompel, 2001). On the other hand, this research contributed with a case study of open access analytics provides an understanding of open access analytics using an OAI-PMH service provider, by demonstrating the processes and unfold challenges and limitations of this approach. This is in contrast to the existing literature, which provides an account of OAI-PMH harvesting and data processing in light of information services (Liu et al., 2005). In addition, it provides an empirical contribution, by reporting ROAR as a contemporary case study that has suffered from unrecoverable failure. The explanatory case study in Chapter 4 demonstrates the limitations of open access registries to enable single point of discovery due to the quality of their records and complexity of open access repositories taxonomy (see Section 4.4.1), the complexity of operationalising the unit of analysis in a particular analytics due to the limitations in the OAI-PMH metadata schemes (see Section 4.4.2), the complexity and resources intensive harvesting process due to large volume of data and low quality of OAI-PMH standards adoptions (see Section 4.4.3) and the issue of service provider sustainability due to single point of failure (see Section 4.5).

Second, the doctoral contributed to literature that conceptualises the open access repositories implementations, data analytics process and social machine by introducing the notion open access analytics (see Section 2.1.3) and original conceptual framework denoted as Open Access Analytics using Open Access Repository Data with a Social Machine (OAA-OARD-SM) provided in Chapter 5. OAA-OARD-SM re-conceptualises the open access analytics process as social machine incorporates open access repository infrastructure, open access registries and web observatory. The framework identifies the open access repositories as managed entity drive the open access analytics process through incorporation of the web observatory infrastructure as mediated layer to form social machinery process realising the open access analytics agenda in collective manner. While the existing literature extends the concept of social machines to varies domains (Ahlers et al., 2016; Murray-Rust et al., 2015a; Arafat et al., 2019; Tiropanis et al., 2014b), this research extends the social machines concept to open access analytics and open access repository domains.

Third, this research contributes to the literature of open access repositories by identifying and theming the analytical application within the open access repository organisational boundary. Thus, it provides an understanding of the analytics practices within open access repositories, which shows that a wide form of analytics applications is operated classifying them into two main categories; the distributed analytical applications and locally operated analytics (see Section 6.3.1). The distributed analytics application includes cross-repository OAI based analytics, cross-repository usage data aggregators, solo-repository content centric analytics and solo-repository centric analytics. On the other hand, the locally operated analytics take forms of CRIS, repository embedded functionalities and in-house developed analytics. The qualitative study provided in Chapter 6 also contributes four themes classifies the repository management interactions with analytics into four roles including data analyst role, administrative role, data and system management role, and system development and support role (see Section 6.3.2). Furthermore, it raises concerns associated with the application of analytics on open access repositories including data-related, cost related and analytical concerns (see Section 6.3.3).

7.3 Implications and Recommendations

This work has a number of implications for the open access community as a whole, as well as open access researchers, repository managers and open access analytics service providers. These implications and recommendations are as follows:

- The implication for service providers is the fact it draws focus on reconsideration of the bringing the mutually beneficial arrangement between them and open access repositories. While the information services and scholarly catalogue and databases are based on the support of the institutions' research findings and dissemination of the research output, this may not be the case with pure analytics service providers. Therefore, part of the implication of this research is for service providers to investigate a new form of arrangements that mutually benefit open access repositories and the service provider, thus motivating open access repositories to engage in their services.

- This research has emphasised and supported the guiding principles of next generation repositories, as highlighted in the "*Next Generation Repositories: Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group*" report, which is emphasised on the support of close collaboration and engagement between open access repositories and external service providers. It has also highlighted the web observatories can act as a mediated layer that support the realisation of open access analytics and facilitates two-way communication between open access repositories and service providers. Therefore, it has brought attention to practices contributed by the web science research community that can also support the open access repository community. In addition, it has called on the web observatory research community to consider the open access repository community as potential participants to support their vision.

- By bringing up the notion of 'open access analytics', this research has highlighted the practices that can support open access advocacy and provide insight into open access development and adoption.

- It has given insight that open access analytics practices need to extend region wide and nationwide collaboration to a global scale of collaboration using web technologies and infrastructure.

- The open access repository research community have focused on shaping repository management roles and investigated the shift in the librarian's role and its correlation with institutional repository management (Chan et al., 2005; Allard et al., 2005; Walters, 2007; Wickham, 2010). This effort needs to be continued, with particular attention given to the rise in new relationships with new entities that have emerged within repository institutions and support the analytics practices. However, repository management roles, skills and responsibilities require further investigation, in order to support the emerging practices.

7.4 Research Limitations

The concept of social machines in the literature is mostly limited to a number of well-known examples of web-based socio-technical systems, including Wikipedia, social media, citizen science, and so on. In these examples, the fact the social machine has emerged on a large scale has already been demonstrated. This is in contrast to web observatories, which have still not reached the same scale of adoption, making some researchers question the idea that a web observatory can be classified as a social machine (Brown, 2017). Nevertheless, the concept of social machine has been used to demonstrate and support a wide variety of processes and tasks outside these well-known examples of social machines (Murray-Rust et al., 2014). Accordingly, they draw a road-map for these tasks to be realised as a social machine. Similarly, this research has called for the engagement of the open access repository community with web observatories and demonstrated how a web observatory can underpin the open access analytics process. This form of research is important to support the adoption of web observatories and promoting them within research communities other than the web science community.

Furthermore, social machines are an evolving concept influenced by rapid changes in technologies and practices. According to Shadbolt et al. (2019), it is challenging to adopt a particular definition that draws a definite line around what can be considered a particular socio-technical system as a social machine. This leads them to adopt the term social machine in ‘a loose and baggy sense’. This type of challenges makes it difficult to adopt the concept, as there is no consensus on a particular definition. In addition, the treatment of evidence from social machine literature requires careful examination in terms of whether it is consistent with the researcher’s adopted view.

One of the challenges associated with this research was the position of the research in relation to the studied subject. The researcher is an outsider of the open access repository infrastructure. Thus, the demonstration of a particular solution using action-based research requires a level of access to the community. Although the design-based research can be an option, however in this research, the researcher did not aim to design a new solution, but instead consider open access analytics as a social machine that takes advantage of existing infrastructure and systems provided by the web science community. This was done using a number of methods, namely the conceptual framing of open access analytics and open access analytics as a social machine, in order to draw propositions that could be investigated, and assumption drawn from it. Thus, the open access repository management is determined as participants to the process. Hence, the expert interviews were used to reveal the practices at the repository level.

ROAR analytics as service visualise the repository deposit activities were down for a number years, thus making it challenging to study it, as contemporary situation as it is slightly disconnected from its community. This, in turn, led to the research strategies such as the case study being unfruitful approach for carrying out the overall research process. Even though, the case study was considered by its requirements. Also, the attempt made by the researcher to reconstruct the service were considered to enrich the investigation of the research problem using ROAR case study.

Another challenge was the time limitation, as this research was constrained by a limited time frame. Therefore, the research had to adopt a working approach meets the research aims within the PhD time frame. This influenced the whole approach used in this research. A multimethod approach can require more effort, time and resources as a result of pluralism in the research methods. However, this research used multi methods to overcome the challenges associated with this research and enables an effective approach to realise the research aims and answer the research questions.

7.5 Opportunities for Future Research

The OAA-OARDA-SM connects the social machine concept and the fulfilment of open access analytics. Part of social machine ecology is the social machine participants which are presented with a key role in the realisation of social machine process (Smart et al., 2014; Shadbolt et al., 2019). Thus, the social machine research community draws particular attention to understand the participant of the social machine with a number of efforts carried out to on their typology (Ross et al., 2010), personality (ibid), skills (Satzger et al., 2013) and motivations (Zheng et al., 2011).

These efforts are motivated by the rise of heterogeneity in the participants engaged in the crowdworks platforms in terms of their motivation, skills and demographics (Howe, 2009; Ross et al., 2010). Satzger et al. (2013) attributed the challenges to the quality of the output of crowdsourcing platforms to this variety and heterogeneity of participants and lack of manageability. Therefore, a recommendation system is proposed to organise the task assignments and improve the manageability of the crowd workers (Borchert et al., 2017). These efforts are based on the categorisation of the task, which is based on the skills required, as well as the identification of participants based on their skills, with the task, participant selection and recommendations based on these categories.

Thus, the task assignment policy is highly dynamic, based on the capabilities and skills required of the task. On the other hand, a social machine with a role-based task assignment policy is targeted at a specific community. Murray-Rust et al. (2014) proposed a model and techniques to coordinate the software development process between two communities with social machines, in cases where a specific role is assigned to each community member. Their model supports an existing community with their existing roles, as the role of a social machine model is to increase the coordination between the two communities.

Similarly, the OAA-OARD-SM framework positions repository management as a potential participant of the analytics process. However, in the open access repository ecology, repository management is not presented with the data pre-processing role for analytics purposes, in contrast to the communities presented in Murray-Rust et al. (2014) model. While the qualitative study highlighted in Chapter 6 provided an insight into the relationship between the repository management team and analytics practices, there is an opportunity for further research models the skills and the role of the repository management toward the analytics activities on top of their repository data.

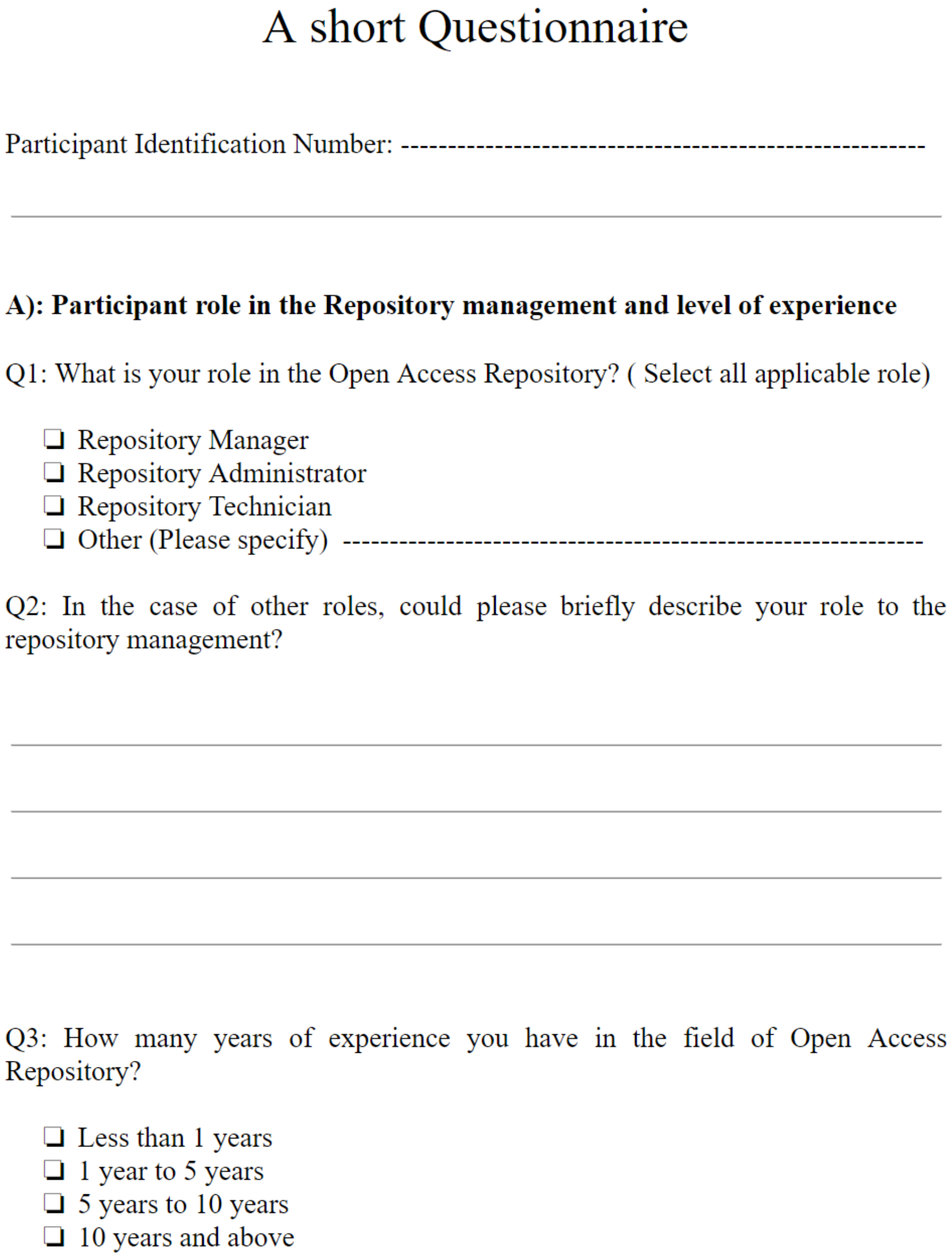
Beside the skills, task complexity is one of the research topics of the social machine research. Yang et al. (2016) investigated the capabilities of crowd workers to perceive task complexity, and attempted to model the task complexity (ibid). They approached task complexity based on the relation between the task and the performer, or what is termed as subjective complexity. Subjective complexity draws a direct connection between the task characteristics and the capabilities of the person working on it. Indeed, a set of factors, which are extrinsic to the task including motivation, facilities and skills, influence the complexity of the task. Therefore, another opportunity for future research is the investigation of the subjective complexity of data analytics tasks for the repository management team.

7.6 Chapter Summary

This Chapter provides a summary of the research aims, contributions, implications and future research directions. In summary, the thesis has discussed the analytics practice of the open access publishing agenda, which supports the monitoring and advocacy of open access publishing. The thesis specifically has drawn attention to the practices that are used with open access repository data as a data source and highlight the delivery approaches of analytics services. As well as this, it has investigated the social machine concept as an approach to realising the analytics process, which directs the inventions of analytics practices at the repository level. This Chapter has also highlighted a set of implications for service providers, researchers and open access repositories. It has also reviewed and emphasised the contributions made by this research, specifically a call for further research into repository management skills and the complexity of analytics in relation to repository management teams.

Appendix A

The Short Questionnaire



A short Questionnaire

Participant Identification Number: ---

A): Participant role in the Repository management and level of experience

Q1: What is your role in the Open Access Repository? (Select all applicable role)

- ❑ Repository Manager
- ❑ Repository Administrator
- ❑ Repository Technician
- ❑ Other (Please specify) --

Q2: In the case of other roles, could please briefly describe your role to the repository management?

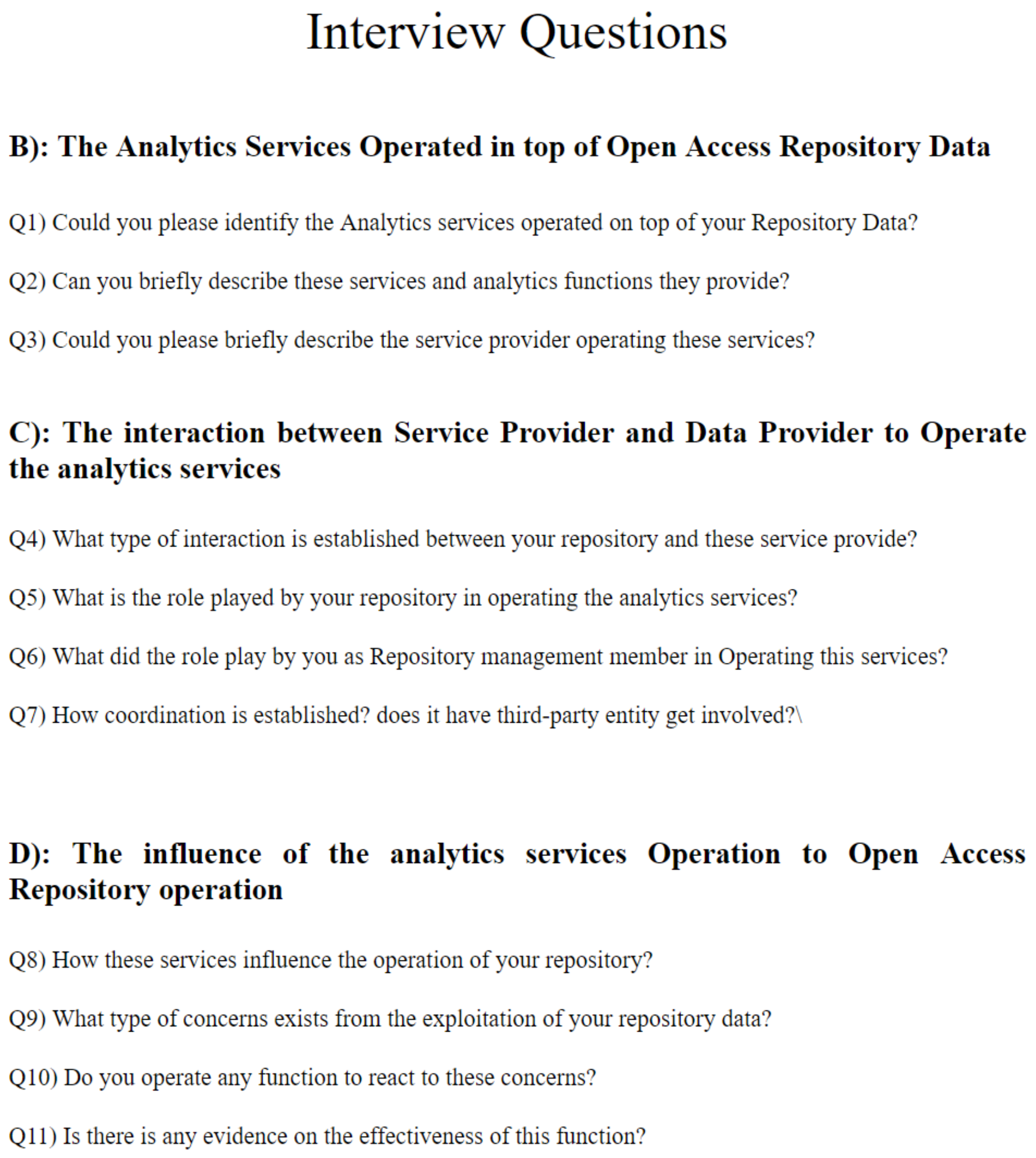
Q3: How many years of experience you have in the field of Open Access Repository?

- ❑ Less than 1 years
- ❑ 1 year to 5 years
- ❑ 5 years to 10 years
- ❑ 10 years and above

FIGURE A.1: The short questionnaire that is designed to collect the participant role and years of expertise

Appendix B

The interview questions



Interview Questions

B): The Analytics Services Operated in top of Open Access Repository Data

Q1) Could you please identify the Analytics services operated on top of your Repository Data?

Q2) Can you briefly describe these services and analytics functions they provide?

Q3) Could you please briefly describe the service provider operating these services?

C): The interaction between Service Provider and Data Provider to Operate the analytics services

Q4) What type of interaction is established between your repository and these service provide?

Q5) What is the role played by your repository in operating the analytics services?

Q6) What did the role play by you as Repository management member in Operating this services?

Q7) How coordination is established? does it have third-party entity get involved?\

D): The influence of the analytics services Operation to Open Access Repository operation

Q8) How these services influence the operation of your repository?

Q9) What type of concerns exists from the exploitation of your repository data?

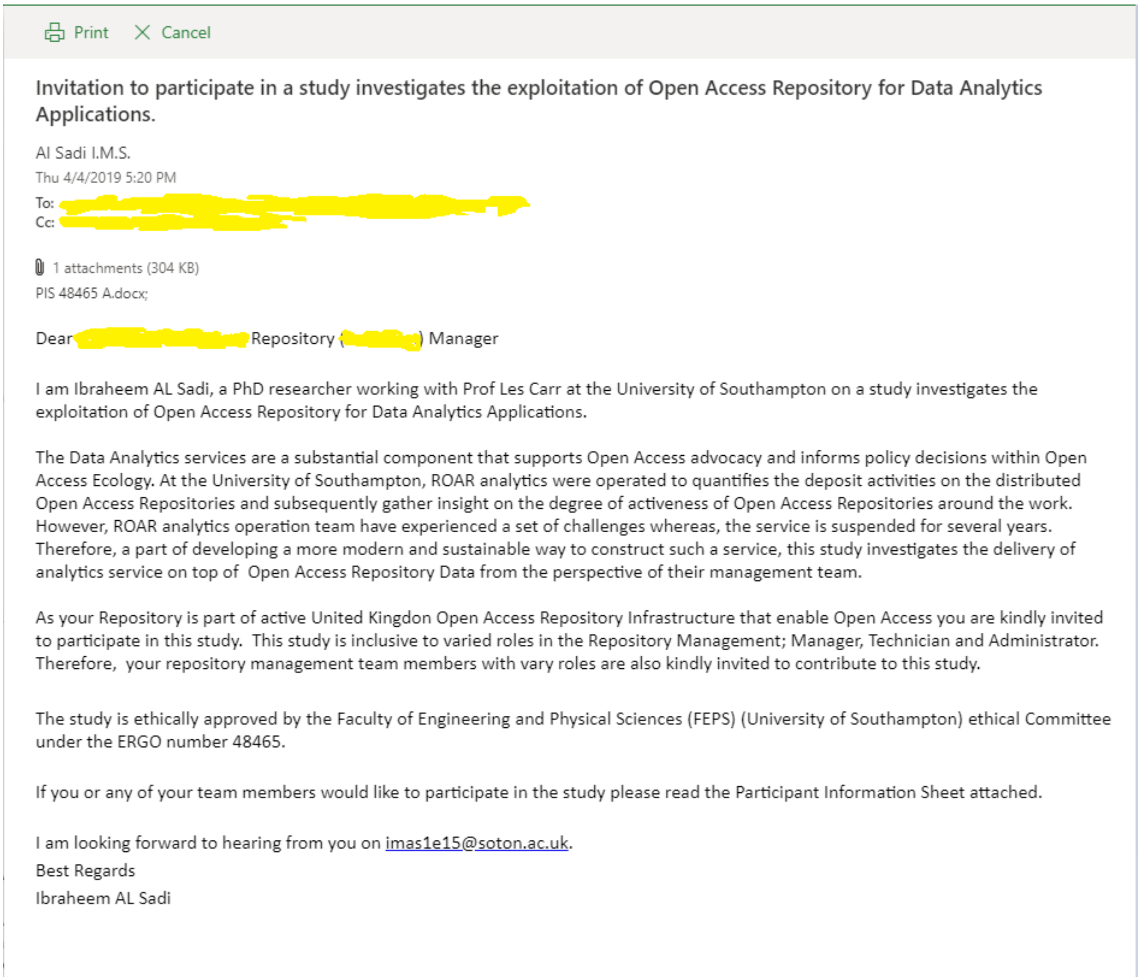
Q10) Do you operate any function to react to these concerns?

Q11) Is there is any evidence on the effectiveness of this function?

FIGURE B.1: The interview questions

Appendix C

The participants invitation email



Invitation to participate in a study investigates the exploitation of Open Access Repository for Data Analytics Applications.

Al Sadi I.M.S.
Thu 4/4/2019 5:20 PM
To: [redacted]
Cc:

1 attachments (304 KB)
PIS 48465 A.docx;

Dear [redacted] Repository ([redacted]) Manager

I am Ibraheem AL Sadi, a PhD researcher working with Prof Les Carr at the University of Southampton on a study investigates the exploitation of Open Access Repository for Data Analytics Applications.

The Data Analytics services are a substantial component that supports Open Access advocacy and informs policy decisions within Open Access Ecology. At the University of Southampton, ROAR analytics were operated to quantifies the deposit activities on the distributed Open Access Repositories and subsequently gather insight on the degree of activeness of Open Access Repositories around the work. However, ROAR analytics operation team have experienced a set of challenges whereas, the service is suspended for several years. Therefore, a part of developing a more modern and sustainable way to construct such a service, this study investigates the delivery of analytics service on top of Open Access Repository Data from the perspective of their management team.

As your Repository is part of active United Kingdon Open Access Repository Infrastructure that enable Open Access you are kindly invited to participate in this study. This study is inclusive to varied roles in the Repository Management; Manager, Technician and Administrator. Therefore, your repository management team members with vary roles are also kindly invited to contribute to this study.

The study is ethically approved by the Faculty of Engineering and Physical Sciences (FEPS) (University of Southampton) ethical Committee under the ERGO number 48465.

If you or any of your team members would like to participate in the study please read the Participant Information Sheet attached.

I am looking forward to hearing from you on imas1e15@soton.ac.uk.
Best Regards
Ibraheem AL Sadi

FIGURE C.1: The participants invitation email.

Appendix D

Analytics services/tools used by RM team or feed by the repositories

Name	Description	Involve OARD Exploitation	Open Access Analytics
Altmetric Explorer for Institutions	Altemtric based service enables the repository managers and the decision makers within a particular institution to monitor the online activity surrounding their academic research including engagement with research in social media and news, and the research readership in Mandalay.	Yes	No
Altmetric Badges	An embedded tool provided by altmetric.com which is integrated with the repository content to provide resource level analytics on the influence and dissemination of a particular resource withing the social network, news and readership	Yes	No
Bielefeld Academic Search Engine (BASE) Statistics	A lightweight statistics quantifies the BASE service provider collection and a chronically illustrate its development.	Yes	Yes
Clarivate with Web of Science Analytics	A powerful platform integrate the Web of Science citation database with a set of traditional citation index and indicators on the global research performance.	No	Yes
Elsevier SciVal	Research performance analytics service target the institutional level analysis. In addition, it provide a set of visualisation tool.	Yes	No
QlikView with Eprints data	In-house Analytics service take the advantage of Eprins data and integrate it with QlikView reporting tool enable a set of analytics provided for the institution academic staff.	Yes	Yes
Google Analytics	A web analytics service operated by Google offers the website owner with a website traffic tracking and reporting services.	No	No

Table D.1: Analytics services/tools used by repository management team or feed by the repositories.

Name	Description	Involve OARD Exploitation	Open Access Analytics
Google Scholar	A global scholarly search engine service equipped with a set of metrics measure the research performance in the global scholarly publishing system	Yes	No
IRStats1 and IRStats2	An Eprints embedded analytics service provide a statistical overview on the repository content of, and the usage and download statistics.	Yes	Yes
IRUS-UK	a national-wide usage statistics aggregation service, that provides a usage statistics for UK institutional repositories.	Yes	Yes
COnnecting REpositories (CORE)	An open access repository aggregator harvest all forms of open access content distributed across different systems worldwide, to enrich it and provide added value services including analytical services	Yes	Yes
Jisc Publications Router	A Jisc service enable the UK Institutions to capture the its affiliated publication from the content providers such as publishers which enables the repository managers to examine the Open Access compliance's within their institution using analytics tools.	Yes	Yes
OpenAIRE	a European Open Science project establish an open and sustainable scholarly communication infrastructure support the overall management, analysis, manipulation, provision, monitoring and cross-linking of all research outcomes.	Yes	Yes
Power BI	Analytics tools provides interactive visualisation and business intelligent capabilities. [a]	Yes	-

TABLE D.2: Analytics services/tools used by repository management team or feed by the repositories (Table Continued).

[a] This is a general purpose analytics tool can work for wide scenario and agenda, P(01) report a general usage of it to support a set of manual analytics activities of the repository data.

Name	Description	Involve OARD Exploitation	Open Access Analytics
Current Research Information System (Pure and Symplectic Elements)	an information system store, manage and organise metadata about the institution research powered with analytcis and reporting capabilities	Yes	Yes
Researchfish	A analytics platform for research funder and institution to track their funding and report the impact of it.	Yes	No
Scopus	A global commercial scholarly database index the research output provides wide forms of analytics services.	Yes	No
Zabbix	A network and infrastructure monitoring tool	No	No
Matomo	A web analytics platform enable a analytics on repository web pages views and resource downloads.	No	No

TABLE D.3: Analytics services/tools used by repository management team or feed by the repositories (Table Continued).

Appendix E

RM Interactions Themes with the Analytics Applications

Main Theme	Sub-theme	Illustrative quotes
Data Analyst	Analyse adherence to open access policy	”...,the availability of open access items, for example, within the repository. Whether they adhere to the REF open access policy, which is a big thing at the moment.” P(01)
		”sometimes [..] the data that in pure and therefore that in prints is sometimes incomplete because it realised on our authors to upload it. so some of the places to look on is Scopus a big abstracting and indexing services ... what we do run jobs where we pull new content that has our [the institution] authors on it from Scopus into pure which then populate it to the eprints ... in terms of analytics .. we sometimes looks at scopues .. dawnload usually as CSV a list of papers say for instance, if we doing a return to [National Institute for Health Research] NIHR where we saying we want to look at the last three years worth of content published as result of NIHR work and whether that is open access. So what might do we got a list of authors associated with NIHR write a billion logic search in Scopus to find all their content or look only of the [the health related research centre in the institution] or something like that. and then we pull that dawn but the thing is we left with then information such this DOI title authors. so it is not really very help full scoupes in term of looking on open access. so what we then do maybe run a search in a service called Unpaywall which integrate some of this information drawn from eprints.... but basically you can use that to tell you some interesting information to inform decision that you make ... but if you really want a good quality information about what gold open access? what green open access? where people are publishing? it is easier to get it from unpaywall by doing a search on the DOI, so unpaywall searcher all the institutional repository and uses cross ref and other services to pull togather massive data and information .. so if you send them a list of DOI that you what to look at and they will send you back a list of doi with information on whether it is open access and what sort of open access it is. so what then you can do in excel is apply filters or use private table to basically say is the journal is fully open access yes or no .. and than you can say is the article is open access and if ti is.” P(10)
	Analyse community and user behaviours	”..they can very clearly indicate that these are the downloads, we think, from real people.” P(10)

Table E.1: Repository Management interaction themes with the analytics applications

Main Theme	Sub-theme	Illustrative quotes
		”It [analytics service] provides a mix of downloads and also access. So, both of those things aren’t the same. So, it’s really interesting to see where the interest is versus where the downloads are.” P(05)
		”we use that [altmetric], so that gives us some analytics in terms of attention around content that we surface in the repository through the digital object identifier.” P(05)
		”it [the analytics service] would give us fairly robust evidence that we can rely on, because it has stripped all of these bots and harvesting and so on, out. So, I think it’s very useful at that individual level but also at the aggregate level of the repository. So, we did some numbers in our… We looked at just over a million downloads for 2017/18 for our publications repository, but I actually think, thinking about, with open access theses can actually be really important as well. Now, a lot of places will probably combine their postgraduate theses with their research outputs. So, for 2017/18 we had about 750,00 downloads for our theses, but the actual number of theses that we have is quite small compared to the number of papers that are available in the publications repository. So, there is very clear open access evidence of interest in the content of the theses that we make available.” P(05)
	Benchmark the repository with other repositories	”you can go onto the IRIS UK website and you can compare yourself against other organisations.” P(02)
	Examine the repository content cross-repository distribution	”It’s starting to look at ORCIDs which are embedded in repositories and so on as well, so there’s a very interesting analytics piece around the data which IRUS-UK is capturing that at a national level you can then look at to go, ah right, yes, this DOI appears in however many different repositories. Or someone with this ORCID has content across a multiple number of repositories and what do the downloads and the related statistics that are aggregated around that look like?” P(05)
		”It [the analytics platform] will allow you to export it as a CSV and manipulate it as you like, to produce statistics and for other insights, which the top universities like being able to benchmark themselves against each other” P(07)
	Communicate Analytical Insight	”We can also run internal reports to feedback to colleges about various things, such as issues around items that may not be compliant for the next ref exercise. So, we can run internal reports like that, as well.” P(05)

Table E.2: Repository Management interaction themes with the analytics applications (Table Continued).

Main Theme	Sub-theme	Illustrative quotes
		”And that’s the group that we provide reports to, it’s every two months, to tell them how is compliance going with the ref policy here at the university and if there’s any issues with that. But we use QlikView for that because it’s got the nice visuals.” P(05)
		”to inform funding bodies and national pressure group” P(10)
Administrative Role	Metadata De-Duplication	”There are a few minor side-effects, because it automatically goes into the repository and we could probably technically fix it, but sometimes we end up with duplicates that need to be de-duplicated” P(03)
	Content Management	”So, institutions can have a dashboard which helps them manage the content that’s harvested into Core, but Core have also been looking at ways in which they’re able to bring together some more analytics data based on the publications set, which they can pull out.” P()
		”we can go into Core if we want, perhaps to take a document down, that’s been harvested by them. So, we have that kind of interaction” P(05)
		”process manuscripts that are deposited via Symplectic, and then come into us, and we check the metadata” P(06)
		”Basically, anything that goes live into our system has to be checked by a human.” P(06)
		”... full text deposit is managed by us in the open access team. So, anyone uploads something, we have to process and manage it, and we have to either accept or reject. And often, we have to have conversations with academics, you’ve uploaded the wrong version. You know, if someone uploads an entire book, we need to say, do you have the copyright permission to upload this?” P(06)
	Integrate External metadata with analytics system metadata	”the metadata with those data sources [eternal metadata sources] will merge with the accepted manuscript. So, the accepted manuscript will have a title, an abstract, but it won’t necessarily have page numbers. It won’t have issue date. Maybe doesn’t have a DOI.” P(06)
	Metadata Quality Control	”I work with the day to day operations team, so we are responsible for running the [unclear], so we get the best metadata, full text up that we can, which of course informs the various [unclear] from that.” P(04)

Table E.3: Repository Management interaction themes with the analytics applications (Table Continued)

Main Theme	Sub-theme	Illustrative quotes
Data and System Management	Manage the policy side	"I would say we manage the policy side of it. We manage the processing. We respond to authors and users, external users." P(06)
	Liaison the data with external groups	"external liaison the data with external groups" P(10)
		"They're not necessarily providing us with a service; we are definitely providing them with statistics as are dozens and dozens and dozens of other universities." P(02)
		"I don't get provided access to the information I send them, but I can see what they do with that information. So it's all high-level numbers," P(02)
		"one of the things that we probably could be doing more of is, yes, looking at the Core dashboard, potentially engaging a bit more periodically with OpenDOAR, which we need to curate and manage ourselves." P(05)
	Liaison the data with internal groups	"[the] team helps people to get data off of [the repository], maybe individuals want to know can I have a list of my publications or they want to have a list of someone else's publications, then the team will help people do that" P(03)
		"We also produce the data for, we have what we call a research, planning and strategy committee here." P(03) "We also report on creation of all types of content for academic users for performance enhancement reviews for certain schools. So, yes, it's a variety of purposes that the reports are for." P(01)
	Coordinate user groups	"....facilitate or coordinate the user groups in the university within the three collages" P(01)
		"What we would do then is to encourage.... Because researchers have to upload or to sign off on the submissions themselves on the Researchfish." P(01)

Table E.4: Repository Management interaction themes with the analytics applications (Table Continued)

Main Theme	Sub-theme	Illustrative quotes
System Development and Support	Add New Functionalities	”So, my team is responsible for open access, so if there’s ay requirements or new add-ons to the service, whether they are technical or otherwise, then we’ll decided whether or not those need to go in. And usually we’ll resource putting them in and setting them up and making sure it’s integrated with the technical side and also the process side of things.” P(03)
		”Yes, he’s [the technician] in a different team but he’s part of the local support that we have around that. So, when we’re creating new services or adding new item types or we’re working with him around the issue with connecting to IRUS-UK. He would be involved in the migration that we’re doing in order to sort, in order to... One of the reasons our IR Stats Two isn’t working at the moment is we need to reconfigure of NODB backend table in MySQL and we’re going to do that as part of a migration to a virtual machine environment”P(04)
		”Or also, dealing with any queries, so for instance we have had queries if statistics aren’t available, what statistics we can do, some local work to create some analytics that we can provide on an individual basis.” P(05)
	Check system compliance and analyse system requirements	”individual colleagues sometimes might look at these things but generally speaking we don’t use them in my team. They’re just there, we’re just giving the information to them. So, we’re just ticking the compliance box. ” P(03)

Table E.5: Repository Management interaction themes with the analytics applications (Table Continued)

Main Theme	Sub-theme	Illustrative quotes
		”it came through as a statement from the EU saying that you should be OpenAIRE compliant if you receive funding from them. And we checked our repository and we felt it was, but that we would just, we just mapped the fields[?] we had and then went back to them and they checked and said, yes, you are compliant. So, yes, it’s just there now, we don’t really have any ongoing interaction other than hearing occasional news items about developments and that aspect. So, it works, it’s there. IV Yes, so if you check your compliance completely, it is your responsibility to check that? IE Yes.” P(03)
		”No, that’s managed by the IRUS side of things. We have to make sure it’s compliant, but once the material is available and the data is available, then they do the processing on their end.” P(07)
		”Because it hasn’t been upgraded for many years, we needed to upgrade so many levels so that it can meet the technological requirements of Symplectic, which is Repository Tools 2. And so, that means they can feed Symplectic into DSpace. So, yes, lots of upgrades. We, essentially, have to communicate with Symplectic to say, do this. Yes.” P(06)
		”It seems to harvest not just our metadata, but also the content in our PDFs. So, there’s been questions around those sorts of things, whether our content needs to be enhanced or changed to fit with those repository aggregator services” P(06)
	Join Pilot Project and Provide feedback	”Jisc were doing a pilot of this and they asked us if we would like to be involved. So, we were involved quite early on, maybe even the first site, I’m not sure, that did this. And they just said, are you interested in doing this, we said yes.” P(03)
		”So they’ve done a bit of work on how the statistics look. And they have sought feedback. I haven’t fed back to them, but they have at least been proactive and come to us and said, right, so we can now do this; what do you guys think?” P(02)

Table E.6: Repository Management interaction themes with the analytics applications (Table Continued)

Main Theme	Sub-theme	Illustrative quotes
	Technical troubleshooting	”I guess we would liaise with the vendor or with Altmetric if there was an issue around that in terms of harvesting the data or perhaps not seeing Altmetric data where we would expect to see it.” P(05)
		”So there was a bit of interaction there, but that was just a case of us trying to fix it at our end. They weren't necessarily coming to us and talking about the sorts of reports that they could do.” P(02)
		”The only interaction I've had is fixing one of our feeds, one of this feeds, the thing that goes to IRIS UK.” P(02)
		”So the role I had there was to work with the IRIS UK to work out why our stuff wasn't making it to their servers. So there was trouble-shooting at my end and trouble-shooting at their end,...”P(02)
		”the work we're doing at the moment with IRUS-UK and the plug in, the troubleshooting is between the technical team at IRUS-UK, our own server team and our local developer here, in order to, so, look at the log files, resolve the issue that has arisen.” P(05)
		”the team I work with has the easier job of saying, this is broken, can you fix it? (the interviewer) So, you mean the service provider himself will troubleshoot it? Yes, we'll work with them around that. But if it is from you end, in terms of the configuration is not well reformed or something? So, that is with myself and our local developer and our local server team here. So, if we think about. . . ” P(04)
		”We needed to contact them. So, I think we needed to provide them with an updated OAI-PMH. So, that was a data issue, in that they weren't retrieving as much as they should have, and there was a. . . It hadn't been done for a few months, so they were missing out a lot of our content. I mean, we publish 12,000-a-year publication. So, anything that they haven't done for six months, they're missing out a lot of our content.” (P06)

Table E.7: Repository Management interaction themes with the analytics applications (Table Continued)

Main Theme	Sub-theme	Illustrative quotes
		”If there are any problems that we have, that we notice—for example, if content is unavailable, or if the website is down, or if there are problems with us processing—we go to them.” P(06)
		”We tried to figure out… I think we discovered by accident why they hadn’t been harvesting our repository as regularly as they… And so, they asked us to… They would do another technical fix to make sure that it was… And so, there was a problem with that, but that was… It was not something that we were aware of.” P(06)
		”And, yes, so I think I must have done some testing to make sure that they were getting our stuff, and the report that we got back was that it wasn’t technically compliant with EU funding because it wasn’t getting the full text. But as Jisc said to us, they weren’t sure themselves why that was not happening, and that’s the last thing we heard about it.” P(06)
		”My main job in those circumstances, if something isn’t right, my role isn’t to fix it. My role is to liaise with a Technical Lead, to make sure it is fixed, so that I can continue to get the information I need out of it. In that sense, I am the main person through whom all of the statistics and the information gained from the analytics systems…” P(07)
		”I guess we would liaise with the vendor or with Altmetric if there was an issue around that in terms of harvesting the data or perhaps not seeing Altmetric data where we would expect to see it.” P(05)
		”I don’t think it’s so. It would be basically on different aspects that we’ve got collaborations or we’ve got membership of the Open Access Colin [?] Group, for example. We’ve got membership of the UK user group for PURE, for example, but we don’t actually… In terms of managing and administrating of the system there is no nationwide collaboration.” P(01)
		”I have a gentleman on my team who does the programming so he just worked with Jisc to set this up. We don’t have an awful lot of communications, they have a mailing list, so they let us know if there’s any problems or anything comes up or new publishers join the service.” P(03)

Table E.8: Repository Management interaction themes with the analytics applications (Table Continued)

Main Theme	Sub-theme	Illustrative quotes
	Upgrade System	”we do system testing, we try out different configurations and document... Create all the reports and documentation to go with the upgrades.” P(01)
		”Because it hasn’t been upgraded for many years, we needed to upgrade so many levels so that it can meet the technological requirements of Symplectic, which is Repository Tools 2. And so, that means they can feed Symplectic into DSpace. So, yes, lots of upgrades. We, essentially, have to communicate with Symplectic to say, do this. Yes.” P(06)
		”Yes, he’s in a different team but he’s part of the local support that we have around that. So, when we’re creating new services or adding new item types or we’re working with him around the issue with connecting to IRUS-UK. He would be involved in the migration that we’re doing in order to sort, in order to... One of the reasons our IR Stats Two isn’t working at the moment is we need to reconfigure of NODB backend table in MySQL and we’re going to do that as part of a migration to a virtual machine environment.” P(05)
		”we interact quite a lot with the service provider Elsevier because they are the proprietors of PURE. In terms of upgrades and for any issues with the system, with the application, we would get in touch with them just to find out if there were any bugs or anything, any issues, any improvements that we can suggest. Because we have quite a large user base and they always come up with ideas for improvements which we pass on to Elsevier.” P(01)
	Monitor system status	”I’ve heard that, and I don’t know if we had any issues with that stuff, apart from the harvesting, irregular harvesting, which seems to have been fixed. I mean, I do check every now and again to make sure that they’re harvesting our most recent material.” P(06)
		”It’s my responsibility to make sure that they’re working, but I can’t actually fix it myself. And it’s my role to then make sure it gets passed on to whoever requires it, needs it, or has asked for it in regard to the data that is produced by the analytics systems.” P(07)

Table E.9: Repository Management interaction themes with the analytics applications (Table Continued)

Bibliography

Abadi, D. J. (2009). Data management in the cloud: Limitations and opportunities. *IEEE Data Engineering Bulletin*, 32(1):3–12.

Abel, R. E., Newlin, L. W., Strauch, K., and Strauch, B. (2002). *Scholarly Publishing: Books, Journals, Publishers, and Libraries in the Twentieth Century.* Wiley. ISBN 978-0-471-21929-3.

Abercrombie, N., Hill, S., and Turner, B. (2006). *The Penguin Dictionary of Sociology.* Credo Reference. Penguin, 5th edition. ISBN 978-0141013756.

Ackoff, R. L. (1989). From data to wisdom. *Journal of applied systems analysis*, 16(1):3–9.

Adler, P. A. and Adler, P. (1994). *Handbook of Qualitative Research*, chapter Observational techniques., page 377–392. Sage publications, Thousand Oaks, California, USA. ISBN 978-0803946798.

Agrawal, D., Das, S., and El Abbadi, A. (2011). Big data and cloud computing: current state and future opportunities. In *Proceedings of the 14th International Conference on Extending Database Technology*, pages 530–533. ACM. doi: 10.1145/1951365.1951432.

Ahlers, D., Driscoll, P., Löfström, E., Krogstie, J., and Wyckmans, A. (2016). Understanding smart cities as social machines. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 759–764. International World Wide Web Conferences Steering Committee. doi: 10.1145/2872518.2890594.

Allard, S., Mack, T. R., and Feltner-Reichert, M. (2005). The librarian's role in institutional repositories: A content analysis of the literature. *Reference services review*, 33(3):325–336. doi: 10.1108/00907320510611357.

Allinson, J. (2006). Oais as a reference model for repositories: an evaluation. Technical report, UKOLN, University of Bath.
URL: *http://eprints.whiterose.ac.uk/3464/*

Allinson, J., Francois, S., and Lewis, S. (2008). Sword: Simple web-service offering repository deposit. Accessed: 05.07.2018.
URL: *http://www.ariadne.ac.uk/issue/54/allinson-et-al/*

Amaratunga, D., Baldry, D., Sarshar, M., and Newton, R. (2002). Quantitative and qualitative research in the built environment: application of "mixed" research approach. *Work Study*, 51(1):17–31. doi: 10.1108/00438020210415488.

Amorim, R. C., Castro, J. A., da Silva, J. R., and Ribeiro, C. (2015). A comparative study of platforms for research data management: Interoperability, metadata capabilities and integration potential. In *New Contributions in Information Systems and Technologies*, pages 101–111. Springer International Publishing. doi: 10.1007/978-3-319-16486-1_10.

Amorim, R. C., Castro, J. A., da Silva, J. R., and Ribeiro, C. (2016). A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, 16(4):851–862. doi: 10.1007/s10209-016-0475-y.

Anguera, M. T., Camerino Foguet, O., Castañer Balcells, M., and Sánchez Algarra, P. (2014). Mixed methods in research into physical activity and sport. *Revista de Psicologia del Deporte*, 23(1):123–130. doi: 10.3389/fpsyg.2020.00578.

Anguera, M. T., Blanco-Villaseñor, A., Losada, J. L., Sánchez-Algarra, P., and Onwuegbuzie, A. J. (2018). Revisiting the difference between mixed methods and multimethods: Is it all in the name? *Quality & Quantity*, 52(6):2757–2770. doi: 10.1007/s11135-018-0700-2.

Arafat, S., Aljohani, N., Abbasi, R., Hussain, A., and Lytras, M. (2019). Connections between e-learning, web science, cognitive computation and social sensing, and their relevance to learning analytics: A preliminary study. *Computers in Human Behavior*, 92:478–486. doi: 10.1016/j.chb.2018.02.026.

Armbruster, C. and Romary, L. (2010). Comparing Repository Types: Challenges and Barriers for Subject-Based Repositories, Research Repositories, National Repository Systems and Institutional Repositories in Serving Scholarly Communication. doi: 10.4018/jdls.2010100104.
URL: *www.igi-global.com/article/comparing-repository-types/48203*

Awre, C. (2006). The technology of open access. In Jacobs, N., editor, *Open Access Key Strategic, Technical and Economic Aspects*, pages 55–62. Elsevier. ISBN 978-1-84334-203-8. doi: 10.1016/b978-1-84334-203-8.50006-6.

Bailey, C. W. (2006). 2 - what is open access? In Jacobs, N., editor, *Open Access Key Strategic, Technical and Economic Aspects*, Chandos Information Professional Series, pages 13 – 26. Chandos Publishing. ISBN 978-1-84334-203-8. doi: 10.1016/B978-1-84334-203-8.50002-9.

Baker, S. E. and Edwards, R. (2012). How many qualitative interviews is enough? expert voices and early career reflections on sampling and cases in qualitative research.

Working paper, National Centre for Research Methods Reviews.
URL: *https://eprints.soton.ac.uk/336913/*

Bardi, A., Castelli, D., and Manghi, P. (2015). Openaire initiative: Providing access, monitoring and contextualizing open access publications. In *IAMSLIC Conference Proceedings 2015*. IAMSLIC.

BASE (n.d). BASE - Bielefeld Academic Search Engine — What is BASE? Accessed: 10.09.2020.
URL: *https://www.base-search.net/about/en/*

BDOA (2003). Berlin declaration on open access to knowledge in the sciences and humanities. Accessed: 11.12.2016.
URL: *https://tinyurl.com/k8xdh7e*

Bekaert, J. and de Sompel, H. V. (2005). A standards-based solution for the accurate transfer of digital assets. *D-Lib Magazine*, 11(06). doi: 10.1045/june2005-bekaert.

Bell, J. and Lewis, S. (2006). Using oai-pmh and mets for exporting metadata and digital objects between repositories. *Program*, 40(3):268–276. doi: 10.1108/00330330610681349.

Benatallah, B., Dumas, M., Fauvet, M.-C., and Rabhi, F. A. (2003). Towards patterns of web services composition. In *Patterns and Skeletons for Parallel and Distributed Computing*, pages 265–296. Springer London. doi: 10.1007/978-1-4471-0097-3_10.

Berners-Lee, T. (2000). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*. Harper Business, San Francisco, 1st edition edition. ISBN 978-0-06-251587-2.

Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., and Weitzner, D. J. (2006). Creating a science of the web. *Science*, 313(5788):769–771. ISSN 00368075, 10959203.
URL: *http://www.jstor.org/stable/3846906*

Bhattacherjee, A. (2012). *Social science research: Principles, methods, and practices*. CreateSpace Independent Publishing Platform, University of South Florida, Tampa, Florida, USA, 2nd edition. ISBN 978-1475146127.

Bielik, L., Kosterec, M., and Zouhar, M. (2014). Model metódy (4): Aplikácia a klasifikácia. *Filozofia*, 69(9):737–751.

Björk, B., Roos, A., and Lauri, M. (2009). Scientific journal publishing: Yearly volume and open access availability. *Information Research*, 14(1):paper 391.
URL: *https://tinyurl.com/d3ozpn*

Björk, B.-C. (2005). A lifecycle model of the scientific communication process. *Learned Publishing*, 18(3):165–176. doi: 10.1087/0953151054636129.

Björk, B.-C. (2017). Open access to scientific articles: a review of benefits and challenges. *Internal and Emergency Medicine*, 12(2):247–253. doi: 10.1007/s11739-017-1603-2.

BOAI (2002). Budapest open access initiative - read the budapest open access initiative. Accessed: 2016-12-11.
URL: *https://tinyurl.com/ya3kj6pd*

BOAI10 (2012). Ten years on from the budapest open access initiative: setting the default to open. *JLIS.it*, 3(2). doi: 10.4403/jlis.it-8631.

Bogner, A. and Menz, W. (2009). The theory-generating expert interview: Epistemological interest, forms of knowledge, interaction. In *Interviewing Experts*, pages 43–80. Palgrave Macmillan UK. ISBN 978-0230244276. doi: 10.1057/9780230244276_3.

Bonilla-Calero, A. (2008). Scientometric analysis of a sample of physics-related research output held in the institutional repository strathprints (2000-2005). *Library Review*, 57(9):700–721. doi: 10.1108/00242530810911815.

Borchert, K., Hirth, M., Schnitzer, S., and Rensing, C. (2017). Impact of task recommendation systems in crowdsourcing platforms. In *FATREC Workshop on Responsible Recommendation*, pages 20–25. doi: 10.18122/b2cx1q.

Brace, J. (2008). Versioning in repositories: Implementing best practice. Accessed: 01.05.2018.
URL: *http://www.ariadne.ac.uk/issue/56/brace/*

Bradley, K. (2006). Digital sustainability and digital repositories. In *VALA2006 Proceedings*. Melbourne, Australia. Accessed: 15.06.2018.
URL: *https://tinyurl.com/y6qdzyd3*

Braun, V. and Clarke, V. (2012). Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, APA handbooks in psychology®., pages 57–71. American Psychological Association, Washington, DC, US. ISBN 978-1-43381-005-3. doi: 10.1037/13620-004.

Braun, V., Clarke, V., Hayfield, N., and Terry, G. (2018). *Thematic Analysis*, pages 1–18. Springer Singapore, Singapore. ISBN 978-981-10-2779-6. doi: 10.1007/978-981-10-2779-6_103-1.

Breen, L. (2007). The researcher'in the middle': Negotiating the insider/outsider dichotomy. *The Australian Community Psychologist*, 19(1):163–174. ISSN 1320-7741.

Brody, T. (2003). Citebase search: Autonomous citation database for e-print archives. In *Third international technical workshop and conference of the project SINN*. Oldenburg, Germany. 16 - 18 Sep 2003.
URL: *https://eprints.soton.ac.uk/260677/*

Brody, T. (2004). Citation analysis in the open access world. In *Interactive Media International*. Available at https://eprints.soton.ac.uk/260000/.

Brody, T. (2006). *Evaluating Research Impact through Open Access to Scholarly Communication*. Ph.D. thesis, University of Southampton.
URL: *https://eprints.soton.ac.uk/263313/*

Brody, T., Harnad, S., and Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8):1060–1072. doi: 10.1002/asi.20373.

Brody, T., Carr, L., Gingras, Y., Hajjem, C., Harnad, S., and Swan, A. (2007). Incentivizing the open access research web: publication-archiving, data-archiving and scientometrics. *CTWatch quarterly*, 3(3).

Brown, I. C. (2017). *The DNA of Web Observatories*. Ph.D. thesis, University of Southampton.
URL: *https://eprints.soton.ac.uk/415346/*

Brown, I. C., Hall, W., and Harris, L. (2014). Towards a taxonomy for web observatories. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 1067–1072. ACM, New York,United States. ISBN 978-1-4503-2744-2. doi: 10.1145/2567948.2579212.

Buregio, V., Meira, S., and Rosa, N. (2013). Social machines: a unified paradigm to describe social web-oriented systems. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, pages 885–890. Association for Computing Machinery, New York, NY, USA. ISBN 978-1-4503-2038-2. doi: 10.1145/2487788.2488074.

Buregio, V., Nascimento, L., Rosa, N., and Meira, S. (2014). Personal APIs as an enabler for designing and implementing people as social machines. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 867–872. Association for Computing Machinery, New York, NY, USA. ISBN 978-1-4503-2745-9. doi: 10.1145/2567948.2578829.

Burgin, M. (2006). *Super-Recursive Algorithms*. Monographs in Computer Science. Springer New York. ISBN 9780387268064.

Burégio, V. A., Meira, S. R., Rosa, N. S., and Garcia, V. C. (2013). Moving towards "relationship-aware" applications and services: A social machine-oriented approach. In *2013 17th IEEE International Enterprise Distributed Object Computing Conference Workshops*, pages 43–52. doi: 10.1109/EDOCW.2013.12.

Burégio, V., Brito, K., Rosa, N., Neto, M., Garcia, V., and Meira, S. (2015). Towards Government as a Social Machine. In *Proceedings of the 24th International*

Conference on World Wide Web, WWW '15 Companion, pages 1131–1136. Association for Computing Machinery, New York, NY, USA. ISBN 978-1-4503-3473-0. doi: 10.1145/2740908.2743976.

Burégio, V., Maamar, Z., and Meira, S. (2015). An architecture and guiding framework for the social enterprise. *IEEE Internet Computing*, 19(1):64–68. doi: 10.1109/MIC.2014.85.

Byrne Evans, M., O'Hara, K., Tiropanis, T., and Webber, C. (2013). Crime applications and social machines: crowdsourcing sensitive data. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, pages 891–896. Association for Computing Machinery, New York, NY, USA. ISBN 978-1-4503-2038-2. doi: 10.1145/2487788.2488075.

Callicott, B. B., Scherer, D., and Wesolek, A. (2015). *Making Institutional Repositories Work*. Purdue University Press. ISBN 978-1-61249-423-4.
URL: *https://tinyurl.com/y4gpefuw*

Card, S. K., Mackinlay, J., and Brook, B. S. P. D. S. a. S., editors (1999). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, San Francisco, Calif, 1st edition edition. ISBN 978-1-55860-533-6.

Carr, L. and Brody, T. (2007). Size isn't everything: sustainable repositories as evidenced by sustainable deposit profiles. *D-lib Magazine*, 13(7):4. doi: doi:10.1045/july2007-carr. Accessed: 21.10.2020.

Carr, L., Harnad, S., and Swan, A. (2007). A longitudinal study of the practice of self-archiving. This eprint is being worked up into a journal article by its authors. The accompanying data is also included.
URL: *https://eprints.soton.ac.uk/263906/*

CCSDS (2012). Reference model for an open archival information system (oais). Technical report, CCSDS, Washington DC, USA.
URL: *https://tinyurl.com/yddcgg89*

Centre for Research Communications (2013). The directory of open access repositories - opendoar. Accessed: 27.02.2018.
URL: *http://www.opendoar.org/*

Chan, D. L., Kwok, C. S., and Yip, S. K. (2005). Changing roles of reference librarians: the case of the HKUST Institutional Repository. *Reference Services Review*, 33(3):268–282. ISSN 0090-7324. doi: 10.1108/00907320510611302.

Chang, R., Ziemkiewicz, C., Green, T. M., and Ribarsky, W. (2009). Defining insight for visual analytics. *IEEE Computer Graphics and Applications*, 29(2):14–17. doi: 10.1109/MCG.2009.22.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000). Crisp-dm 1.0: Step-by-step data mining guide. Technical report, CRISP-DM consortium.
URL: *https://tinyurl.com/yyuxg32k*

Chassang, G. (2017). The impact of the EU general data protection regulation on scientific research. *ecancermedicalscience*, 11. doi: 10.3332/ecancer.2017.709.

Chaudhuri, S. and Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM SIGMOD Record*, 26(1):65–74. ISSN 0163-5808. doi: 10.1145/248603.248616.

Chopra, A. K. and Singh, M. P. (2016). From Social Machines to Social Protocols: Software Engineering Foundations for Sociotechnical Systems. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 903–914. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE. ISBN 978-1-4503-4143-1. doi: 10.1145/2872427.2883018.

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1):7–19. ISSN 00032638, 14678284.
URL: *https://tinyurl.com/y5o29r5x*

Conway, P. (2008). Modeling the digital content landscape in universities. *Library Hi Tech*, 26(3):342–354. doi: 10.1108/07378830810903283.

Cooper, A. (2012a). A brief history of analytics. *CETIS Analytics Series*, 1(9):2–21. ISSN 2051-9214.

Cooper, A. (2012b). A framework of characteristics for analytics. *CETIS Analytics Series*, 1(7):1–17.

Cooper, A. et al. (2012). What is analytics? definition and essential characteristics. *CETIS Analytics Series*, 1(5):1–10.

Cox, J. (1998). The great journals crisis: A complex present, but a collegial future. *Logos*, 9(1):29 – 33. doi: 10.2959/logo.1998.9.1.29.

Creative Commons (n.d). About the licenses. Accessed: 18.09.2020.
URL: *https://creativecommons.org/licenses/*

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications, 3rd edition. ISBN 978-1452228372.

Creswell, J. W. and Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory into practice*, 39(3):124–130.

Creswell, J. W. and Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications, London, UK, 4th edition. ISBN 978-1-5063-3020-4.

Cronin, P., Ryan, F., and Coughlan, M. (2008). Undertaking a literature review: a step-by-step approach. *British journal of nursing*, 17(1):38–43. doi: 10.12968/bjon.2008.17.1.28059.

Crow, R. (2002). The case for institutional repositories: A sparc position paper. Technical report, SPARC, Washington, USA.
URL: *https://tinyurl.com/y463d7l2*

Crowe, S., Cresswell, K., Robertson, A., Huby, G., Avery, A., and Sheikh, A. (2011). The case study approach. *BMC medical research methodology*, 11(1):100. doi: 10.1186/1471-2288-11-100.

Daoutis, C. and Rodriguez-Marquez, M. (2018). Library-mediated deposit: A gift to researchers or a curse on open access? reflections from the case of surrey. *Publications*, 6(2). doi: 10.3390/publications6020020.

Dasu, T. and Johnson, T. (2003). *Exploratory data mining and data cleaning*. John Wiley & Sons, Hoboken, New Jersey, USA. ISBN 0-471-26851-8.

Davenport, T. H. et al. (2006). Competing on analytics. *harvard business review*, 84(1):98.
URL: *https://hbr.org/2006/01/competing-on-analytics*

Davenport Thomas, H. and Harris, J. G. (2007). *Competing on analytics: the new science of winning*. Harvard Business School Press. ISBN 978-1422103326.

De Castro, P. (2014). 7 things you should know about institutional repositories, cris systems, and their interoperability. Accessed: 020.08.2018.
URL: *https://tinyurl.com/y7t2xajl*

De Castro, P. (2019). The role of current research information systems (cris) in supporting open science implementation: the case of strathclyde. *ITLib*, 2018:21–30. doi: 10.25610/itlib-2018-0003.

De Silva, P. U. and Vance, C. K. (2017). On the road to unrestricted access to scientific information: The open access movement. In *Scientific Scholarly Communication*, pages 25–40. Springer. doi: 10.1007/978-3-319-50627-2_3.

Delen, D. and Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems*, 55(1):359 – 363. ISSN 0167-9236. doi: 10.1016/j.dss.2012.05.044.

DiFranzo, D., Erickson, J. S., Gloria, M. J. K. T., Luciano, J. S., McGuinness, D. L., and Hendler, J. (2014). The web observatory extension: Facilitating web science collaboration through semantic markup. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 475–480. ACM. doi: 10.1145/2567948.2576936.

Dodds, F. (2018). The changing copyright landscape in academic publishing. *Learned Publishing*, 31(3):270–275. doi: 10.1002/leap.1157.

Doyle, L., Brady, A.-M., and Byrne, G. (2016). An overview of mixed methods research–revisited. *Journal of research in nursing*, 21(8):623–635. doi: 10.1177/1744987116674257.

Eckert, C. and Stacey, M. (2010). What is a process model? reflections on the epistemology of design process models. In *Modelling and Management of Engineering Processes*, pages 3–14. Springer. ISBN 978-1-84996-198-1.

Eprints Group (2004). The registry of open access repository (roar). Accessed: 27.02.2018.
URL: *http://roar.eprints.org/*

Eprints Group (n.d). Oai-pmh registered data providers. Accessed: 15.05.2018.
URL: *https://www.openarchives.org/Register/BrowseSites*

Eve, M. P. (2014a). *Digital economics*, page 43–85. Cambridge University Press. doi: 10.1017/CBO9781316161012.004.

Eve, M. P. (2014b). *Introduction, or why open access?*, page 152–178. Cambridge University Press. doi: 10.1017/CBO9781316161012.008.

Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS biology*, 4(5):e157. doi: 10.1371/journal.pbio.0040157.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., and Pappas, G. (2008). Comparison of pubmed, scopus, web of science, and google scholar: strengths and weaknesses. *The FASEB journal*, 22(2):338–342. doi: 10.1096/fj.07-9492LSF.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54. doi: 10.1609/aimag.v17i3.1230.

Fecher, B. and Friesike, S. (2014). Open science: one term, five schools of thought. In *Opening science*, pages 17–47. Springer. doi: 10.1007/978-3-319-00026-8_2.

Federico, P., Heimerl, F., Koch, S., and Miksch, S. (2017). A survey on visual approaches for analyzing scientific literature and patents. *IEEE transactions on visualization and computer graphics*, 23(9):2179–2198. doi: 10.1109/TVCG.2016.2610422.

Ferros, L., Ramalho, J. C., and Ferreira, M. (2008). Creating a national federation of archives using oai-pmh. In *XATA2008 XML: Applications and Associated Technologies 6th National Conference*, pages 12–21. XML: Applications and Associated Technologies.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, 12(2):219–245. doi: 10.1177/1077800405284363.

Foddy, W. and Foddy, W. H. (1994). *Constructing questions for interviews and questionnaires: Theory and practice in social research*. Cambridge university press, Cambridge, UK. ISBN 0-521-46733-0.

Foster, N. F. and Gibbons, S. (2005). Understanding faculty to improve content recruitment for institutional repositories. *D-Lib Magazine*, 11(1). doi: 10.1045/january2005-foster.

Freeman, R. E. (1984). *Strategic management: A stakeholder approach*. Cambridge University Press, Cambridge, UK. ISBN 978-0-521-15174-0.

Froschauer, U. and Lueger, M. (2009). Expert interviews in interpretive organizational research. In *Interviewing experts*, pages 217–234. Springer. ISBN 978-0-230-24427-6.

Gadd, E., Oppenheim, C., and Probets, S. (2003a). Romeo studies 1: The impact of copyright ownership on academic author self-archiving. *Journal of documentation*, 59(3):243–277. doi: 10.1108/00220410310698239.

Gadd, E., Oppenheim, C., and Probets, S. (2003b). Romeo studies 2: How academics want to protect their open-access research papers. *Journal of information science*, 29(5):333–356. doi: doi.org/10.1177/01655515030295002.

Gadd, E., Oppenheim, C., and Probets, S. (2004). Romeo studies 5: Ipr issues facing oai data and service providers. *The Electronic Library*, 22(2):121–138. doi: 10.1108/02640470410699143.

Gallen, C. (2013). Some considerations for a web observatory. In *1st International workshop on Building Web Observatories*.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., and Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, 1(1):9. doi: 10.1186/s41044-016-0014-0.

Genoni, P. (2004). Content in institutional repositories: a collection management issue. *Library management*, 25(6/7):300–306. doi: 10.1108/01435120410547968.

Ghosh, M. (2011). Advocacy for open access: a selected review of the literature and resource list. *Library Hi Tech News*, 28(2):19–23. doi: 10.1108/07419051111135245.

Gill, P., Stewart, K., Treasure, E., and Chadwick, B. (2008). Methods of data collection in qualitative research: interviews and focus groups. *British dental journal*, 204(6):291–295. doi: 10.1038/bdj.2008.192.

Gläser, J. and Laudel, G. (2009). On interviewing "good" and "bad" experts. In *Interviewing experts*, pages 117–137. Springer. ISBN 978-0-230-24427-6.

Glisczinski, D. (2018). Thematic analysis. *Journal of Transformative Education*, 16(3):175–175. doi: 10.1177/1541344618777367.

Gläser, J. and Laudel, G. (2010). *Experteninterviews und qualitative Inhaltsanalyse: als Instrumente rekonstruierender Untersuchungen*. VS Verlag für Sozialwissenschaften, Springer, Fachmedien Wiesbaden GmbH, Wiesbaden, 4th edition. ISBN 978-3-531-17238-5.
URL: *https://www.springer.com/de/book/9783531172385*

Goetz, B. (2006). *Java Concurrency in Practice*. Addison-Wesley Professional, Upper Saddle River, NJ. ISBN 978-0-321-34960-6.

Goldkuhl, G. (2012). Pragmatism vs interpretivism in qualitative information systems research. *European journal of information systems*, 21(2):135–146. doi: 10.1057/ejis.2011.54.

Gray, D. E. (2014). *Doing research in the real world*. Sage Publications Ltd, New Delhi, India. ISBN 978-1-4462-6018-0.

Greenberg, A., Hamilton, J., Maltz, D. A., and Patel, P. (2008). The cost of a cloud: research problems in data center networks. *ACM SIGCOMM computer communication review*, 39(1):68–73. doi: 10.1145/1496091.1496103.

Grolemund, G. and Wickham, H. (2014). A cognitive interpretation of data analysis. *International Statistical Review*, 82(2):184–204. doi: 10.1111/insr.12028.

Guédon, J.-C. (2004). The "green" and "gold" roads to open access: The case for mixing and matching. *Serials review*, 30(4):315–328. doi: doi.org/10.1016/j.serrev.2004.09.005.

Guédon, J., Jubb, M., Kramer, B., et al. (2019). Future of scholarly publishing and scholarly communication. Technical report, Corporate author(s): Directorate-General for Research and Innovation (European Commission). doi: 10.2777/836532.

Hacigumus, H., Iyer, B., and Mehrotra, S. (2002). Providing database as a service. In *Proceedings 18th International Conference on Data Engineering*, pages 29–38. doi: 10.1109/ICDE.2002.994695.

Hall, W. and Tiropanis, T. (2012). Web evolution and web science. *Computer Networks*, 56(18):3859–3865. doi: 10.1016/j.comnet.2012.10.004.

Harnad, S. (1994a). Advancing science by self-archiving refereed research. *Science Magazine (Online)*, 1:164–167.

Harnad, S. (1994b). Overture: A subversive proposal. In Okerson, S. and O'Donnell, J. J., editors, *Scholarly Journals at the Crossroads: A Subversive Proposal for Electronic Publishing*, pages 11–12. Association of Research Libraries, Washington, DC.
URL: *https://tinyurl.com/yygdccmc*

Harnad, S. (1995a). The postgutenberg galaxy: how to get there from here. *The Information Society*, 11(4):285–291.

Harnad, S. (1995b). *Scholarly Journals at the Crossroads: A Subversive Proposal for Electronic Publishing*. Association of Research Libraries, Washington, DC. ISBN 0-918006-26-0.

Harnad, S. (2001). The self-archiving initiative. *Nature*, 410(6832):1024–1025. doi: doi.org/10.1038/35074210.

Harnad, S. (2004). For whom the gate tolls? how and why to free the refereed research literature online through author/institution self-archiving, now. *Historical Social Research/Historische Sozialforschung*, pages 76–113.

Harnad, S. (2005). Oa impact advantage = ea + (aa) + (qb) + qa + (ca) + ua. Accessed: 12.04.2017.
URL: *https://eprints.soton.ac.uk/262085/*

Harnad, S. (2006). Publish or perish-self-archive to flourish: the green route to open access. *ERCIM News*, 64. Accessed: 14.06.2018.
URL: *https://tinyurl.com/yypuvm65*

Harnad, S. (2008a). Open access scientometrics and the uk research assessment exercise. *Scientometrics*, 79(1):147–156. doi: 10.1007/s11192-009-0409-z.

Harnad, S. (2008b). Validating research performance metrics against peer rankings. *Ethics in Science and Environmental Politics*, 8:103–107. doi: 10.3354/esep00088.

Harnad, S. and Brody, T. (2004). Comparing the impact of open access (oa) vs. non-oa articles in the same journals. *D-lib Magazine*, 10(6). doi: 10.1045/june2004-harnad.

Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Stamerjohanns, H., and Hilf, E. R. (2004). The access/impact problem and the green and gold roads to open access. *Serials Review*, 30(4):310–314. doi: 10.1016/j.serrev.2004.09.013.

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Hajjem, C., and Hilf, E. R. (2008). The access/impact problem and the green and gold roads to open access: An update. *Serials review*, 34(1):36–40. doi: doi.org/10.1016/j.serrev.2007.12.005.

Hart, C. (1999). *Doing a literature review: Releasing the Social Science Research Imagination*. SAGE Publications, London, UK. ISBN 978-0761959748.

Haw, S. C. and Rao, G. S. V. R. K. (2007). A comparative study and benchmarking on XML parsers. In *The 9th International Conference on Advanced Communication Technology*. IEEE. doi: 10.1109/icact.2007.358364.

Hawkins, B. L. (2008). Accountability, demands for information, and the role of the campus it organization. In Katz, R. N., editor, *The Tower and The Cloud*, pages 98–105. Educause, University of California, Berkeley. ISBN 978-0-9672853-9-9.

Hazen, B. T., Boone, C. A., Ezell, J. D., and Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, 154:72 – 80. ISSN 0925-5273. doi: 10.1016/j.ijpe.2014.04.018.

Hendler, J. and Berners-Lee, T. (2010). From the semantic web to social machines: A research challenge for ai on the world wide web. *Artificial Intelligence*, 174(2):156–161. doi: 10.1016/j.artint.2009.11.010.

Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., and Weitzner, D. (2008). Web science: an interdisciplinary approach to understanding the web. *Communications of the ACM*, 51(7):60–69. doi: 10.1145/1364782.1364798.

Hesse-Biber, S. N. and Johnson, R. B. (2015). *The Oxford handbook of multimethod and mixed methods research inquiry*. Oxford University Press, Madison Avenue, New York, USA. ISBN 978-0-19-993362-4.

Hey, J. M. (2004). Targeting academic research: Southampton's institutional repository. In Lewis, J., editor, *Proceedings of Online Information 2004, 30 Nov-2 Dec 2004*, pages 127–136. Learned Information Europe Ltd.
URL: *https://eprints.soton.ac.uk/13598/*

Hitchcock, S., Bergmark, D., Brody, T., Gutteridge, C., Carr, L., Hall, W., Lagoze, C., and Harnad, S. (2002). Open citation linking. *D-lib Magazine*, 8(10). doi: 10.1045/october2002-hitchcock.

Horwood, L., Sullivan, S., Young, E., and Garner, J. (2004). Oai compliant institutional repositories and the role of library staff. *Library management*, 25(4/5):170–176. doi: 10.1108/01435120410533756.

Houssos, N., Stamatis, K., Banos, V., Kapidakis, S., Garoufallou, E., and Koulouris, A. (2011). Implementing enhanced oai-pmh requirements for europeana. In *International Conference on Theory and Practice of Digital Libraries*, pages 396–407. Springer, Berlin, Heidelberg. ISBN 978-3-642-24469-8.

Howe, J. (2009). *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House. ISBN 978-0307396211.

Hui, S. and Jha, G. (2000). Data mining for customer service support. *Information & Management*, 38(1):1 – 13. ISSN 0378-7206. doi: 10.1016/S0378-7206(00)00051-3.

Hull, R., Neaves, P., and Bedford-Roberts, J. (1997). Towards situated computing. In *Digest of Papers. First International Symposium on Wearable Computers*, pages 146–153. doi: 10.1109/ISWC.1997.629931.

Hutchins, E. (1995). *Cognition in the Wild*. MIT press. ISBN 978-0262082310.

INFORMS (n.d.). Operations Research & Analytics.
URL: *https://tinyurl.com/ybddw99y*

Jabareen, Y. (2009). Building a conceptual framework: philosophy, definitions, and procedure. *International Journal of qualitative methods*, 8(4):49–62. doi: doi.org/10.1177/160940690900800406.

Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Oxford University Press. ISBN 978-0198250616.

Jagadish, H. V., Al-Khalifa, S., Chapman, A., Lakshmanan, L. V., Nierman, A., Paparizos, S., Patel, J. M., Srivastava, D., Wiwatwattana, N., Wu, Y., et al. (2002). Timber: A native xml database. *The VLDB Journal—The International Journal on Very Large Data Bases*, 11(4):274–291. doi: 10.1007/s00778-002-0081-x.

Jantz, R. and Giarlo, M. J. (2005). Digital preservation: Architecture and technology for trusted digital repositories. *Microform & Imaging Review*, 34(3). doi: 10.1515/mfir.2005.135.

Jayathilake, D. (2012). Towards structured log analysis. In *2012 Ninth International Conference on Computer Science and Software Engineering (JCSSE)*, pages 259–264. IEEE. doi: 10.1109/JCSSE.2012.6261962.

Jenkins, C., Probets, S., Oppenheim, C., and Hubbard, B. (2007). Romeo studies 8: Self-archiving: The logic behind the colour-coding used in the copyright knowledge bank. *Program*, 41(2):124–133. doi: 10.1108/00330330710742908.

JISC (2016). Manage your research information | Jisc. Accessed: 20.09.2019.
URL: *https://tinyurl.com/yasmnk4d*

JISC (n.d). About sherpa romeo. Accessed: 26.11.2020.
URL: *https://v2.sherpa.ac.uk/romeo/about.html*

JISC (n.d.a). Monitor uk. Accessed: 21.08.2019.
URL: *https://www.jisc.ac.uk/monitor-uk*

JISC (n.d.b). Supporting open access through metadata and improved interoperability. Accessed: 20.09.2019.
URL: *https://tinyurl.com/y9pszcvt*

Johnson, R. B., Onwuegbuzie, A. J., and Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of mixed methods research*, 1(2):112–133. doi: 10.1177/1558689806298224.

Jones, R. E., Andrew, T., and MacColl, J. (2006). *The institutional repository*. Chandos Publishing, Oxford, England. ISBN 9781780630830.

Jubb, M., Plume, A., Oeben, S., Brammer, L., Johnson, R., Bütün, C., and Pinfield, S. (2017). Monitoring the transition to open access: December 2017. Technical report, The University of Sheffield.

Kallio, H., Pietilä, A.-M., Johnson, M., and Kangasniemi, M. (2016). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of advanced nursing*, 72(12):2954–2965. doi: 10.1111/jan.13031.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., Weaver, C., Lee, B., Brodbeck, D., and Buono, P. (2011). Research directions in data wrangling: Visuatizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288. ISSN 1473-8716. doi: 10.1177/1473871611415994.

Keenan, P. T., Owen, J. H., and Schumacher, K. (2018). Introduction to analytics. *INFORMS Analytics Body of Knowledge*, pages 1–30. doi: 10.1002/9781119505914.ch1.

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G. (2008). *Visual Analytics: Definition, Process, and Challenges*, pages 154–175. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-70956-5. doi: 10.1007/978-3-540-70956-5_7.

Kelly, B., Sheppard, N., Delasalle, J., Dewey, M., Stephens, O., Johnson, G., and Taylor, S. (2012). Open metrics for open repositories. In *OR2012: the 7th International Conference on Open Repositories*.
URL: *http://eprints.leedsbeckett.ac.uk/id/eprint/793/*

Kim, J. (2007). Motivating and impeding factors affecting faculty contribution to institutional repositories. *Journal of digital information*, 8(2):1–11.

Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., and Lee, D. (2003). A taxonomy of dirty data. *Data Mining and Knowledge Discovery*, 7(1):81–99. ISSN 1384-5810. doi: 10.1023/A:1021564703268.

Klein, H. K. and Myers, M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS quarterly*, 23(1):67–94. doi: 10.2307/249410.

Kling, R. (2004). The internet and unrefereed scholarly publishing. *Annual Review of Information Science and Technology (ARIST)*, 38:591–631. doi: 10.1002/aris.1440380113.

Kling, R. and McKim, G. (1999). Scholarly communication and the continuum of electronic publishing. *Journal of the American Society for Information science*, 50(10):890–906. doi: 10.1002/(SICI)1097-4571(1999)50:10¡890::AID-ASI6¿3.0.CO;2-8.

Knoth, P. and Zdrahal, Z. (2012). Core: three access levels to underpin open access. *D-Lib Magazine*, 18(11):4. doi: 10.1045/november2012-knoth.

Kosterec, M. (2016). Methods of conceptual analysis. *Filozofia*, 71(3):220–230.

Kothari, C. R. (2004). *Research methodology: Methods and techniques*. New Age International, New Delhi, India, 2nd edition. ISBN 978-81-224-2488-1.

Koutsomitropoulos, D. A., Tsakou, A. A., Tsolis, D. K., and Papatheodorou, T. S. (2004). Towards the development of a general-purpose digital repository. In *Proceedings of the Sixth International Conference on Enterprise Information Systems - Volume 1: WOSIS*, pages 271–278. ISBN 972-8865-00-7. doi: 10.5220/0002637402710278.

Laakso, M. (2014). Green open access policies of scholarly journal publishers: a study of what, when, and where self-archiving is allowed. *Scientometrics*, 99(2):475–494. doi: doi.org/10.1007/s11192-013-1205-3.

Lagoze, C. and Van de Sompel, H. (2001). The open archives initiative: Building a low-barrier interoperability framework. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '01, page 54–62. Association for Computing Machinery, New York, NY, USA. ISBN 1581133456. doi: 10.1145/379437.379449. **URL:** *https://doi.org/10.1145/379437.379449*

Lagoze, C. and Van de Sompel, H. (2003). The making of the open archives initiative protocol for metadata harvesting. *Library hi tech*, 21(2):118–128. doi: doi.org/10.1108/07378830310479776.

Lavoie, B. F. (2003). The incentives to preserve digital materials: Roles, scenarios, and economic decision-making. Technical report, OCLC Online Computer Library Center, Dublin, Ohio, USA.

Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. (2002). Aimq: a methodology for information quality assessment. *Information & Management*, 40(2):133 – 146. ISSN 0378-7206. doi: 10.1016/S0378-7206(02)00043-5.

Legard, R., Keegan, J., and Ward, K. (2003). In-depth interview. In Jane Ritchie and Jane Lewis, editors, *Qualitative research practice : a guide for social science students and researchers*, chapter 6, page 138. Sage. ISBN 0-7619-7109-2.

Linda, Z. (2018). Open access and scholarly publishing: The scholarly publishing crisis. Accessed: 08.08.2019.
URL: *https://tinyurl.com/yd7pd94h*

Littig, B. (2009). Interviewing the elite – interviewing experts: Is there a difference? In *Interviewing experts*, pages 98–113. Springer. ISBN 978-0-230-24427-6.

Liu, Z. and Stasko, J. (2010). Mental models, visual reasoning and interaction in information visualization: A top-down perspective. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):999–1008. doi: 10.1109/TVCG.2010.177.

Liu, X., Brody, T., Harnad, S., Carr, L., Maly, K., Zubair, M., and Nelson, M. L. (2002). A scalable architecture for harvest-based digital libraries-the odu/southampton experiments. *arXiv preprint cs/0205071*, 8(11). doi: 10.1045/november2002-liu.

Liu, X., Maly, K., Nelson, M. L., and Zubair, M. (2005). Lessons learned with arc, an oai-pmh service provider. *LIBRARY TRENDS*, 53(4):590–603.

Loesch, M. F. (2010). Oaister database http://oaister. worldcat. org. *Technical Services Quarterly*, 27(4):395–396. doi: 10.1080/07317131.2010.501001.

Lossau, N. and Peters, D. (2008). Driver: Building a sustainable infrastructure of european scientific repositories. *LIBER Quarterly*, 18(3-4):437–448. doi: 10.18352/lq.7942.

Lossau, N., Rahmsdorf, S., Pieper, D., and Summann, F. (2006). Bielefeld academic search engine (base) an end-user oriented institutional repository search service. *Library Hi Tech*, 24(4):614–619. doi: 10.1108/07378830610715473.

Luan, H., Hou, D., and Chua, T.-S. (2013). Next-live: A live observatory on social media. In Li, S., El Saddik, A., Wang, M., Mei, T., Sebe, N., Yan, S., Hong, R., and Gurrin, C., editors, *Advances in Multimedia Modeling*, pages 514–516. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-642-35728-2.

Luczak-Roesch, M. and Tinati, R. (2017). The social in the platform trap: Why a microscopic system focus limits the prospect of social machines. Accessed: 012.07.2018.
URL: *https://tinyurl.com/snljpyy*

Luczak-Roesch, M., Tinati, R., O'Hara, K., and Shadbolt, N. (2015). Socio-technical computation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, pages 139–142. doi: doi.org/10.1145/2685553.2698991.

Luczak-Roesch, M., Tinati, R., Aljaloud, S., Hall, W., and Shadbolt, N. (2016). A Universal Socio-Technical Computing Machine. In Bozzon, A., Cudre-Maroux, P., and Pautasso, C., editors, *Web Engineering*, pages 559–562. Springer International Publishing, Cham. ISBN 978-3-319-38791-8. doi: doi.org/10.1007/978-3-319-38791-8_48.

Lynch, C. A. (2003). Institutional repositories: Essential infrastructure for scholarship in the digital age. *Libraries and the Academy*, 3(2):327–336. doi: 10.1353/pla.2003.0039.

Maamar, Z., Mostefaoui, S. K., and Yahyaoui, H. (2005). Toward an agent-based and context-oriented approach for web services composition. *IEEE Transactions on Knowledge and Data Engineering*, 17(5):686–697. doi: 10.1109/TKDE.2005.82.

Maamar, Z., Lahkim, M., Benslimane, D., Thiran, P., and Sattanathan, S. (2007). Web services communities - concepts & operations. In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Volume 2: WEBIST,*, pages 323–327. INSTICC, SciTePress. ISBN 978-972-8865-77-1. doi: 10.5220/0001260103230327.

Mabry, L. (2008). Chapter 13: Case study in social research. In *The SAGE handbook of social research methods*, pages 214–227. Sage Publications. doi: 10.4135/9781446212165.n13.

Madaan, A., Tiropanis, T., Srinivasa, S., and Hall, W. (2016). Observlets: Empowering analytical observations on web observatory. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 775–780. International World Wide Web Conferences Steering Committee. doi: 10.1145/2872518.2890593.

Madill, A., Jordan, A., and Shirley, C. (2000). Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies. *British journal of psychology*, 91(1):1–20. doi: 10.1348/000712600161646.

Manghi, P., Mikulicic, M., Candela, L., Artini, M., and Bardi, A. (2010). General-Purpose Digital Library Content Laboratory Systems. In Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., and Frommholz, I., editors, *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science, pages 14–21. Springer, Berlin, Heidelberg. ISBN 978-3-642-15464-5. doi: 10.1007/978-3-642-15464-5_3.

Mariscal, G., Marban, O., and Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 25(2):137–166. doi: 10.1017/S0269888910000032.

Markey, K., Rieh, S. Y., Jean, B. S., Kim, J., and Yakel, E. (2007). Census of institutional repositories in the united states. Technical report, Council on Library and Information Resources, Washington DC, USA.

Marsh, R. M. (2015). The role of institutional repositories in developing the communication of scholarly research. *OCLC Systems & Services: International digital library perspectives*, 31(4):163–195. doi: 10.1108/OCLC-04-2014-0022.

Martín-Martín, A., Costas, R., van Leeuwen, T., and López-Cózar, E. D. (2018). Evidence of open access of scientific publications in google scholar: A large-scale analysis. *Journal of Informetrics*, 12(3):819–841. doi: 10.1016/j.joi.2018.06.012.

Masinter, L., Berners-Lee, T., and Fielding, R. T. (2005). Uniform resource identifier (uri): Generic syntax. Viewed: 01.05.2018.
URL: *https://tinyurl.com/y3dybbs5*

Mattsson, M. and Bosch, J. (1997). Framework composition: problems, causes and solutions. In *Proceedings of TOOLS USA 97. International Conference on Technology of Object Oriented Systems and Languages*, pages 203–214. doi: 10.1109/TOOLS.1997.654724.

Max-Planck-Gesellschaft (n.d). Berlin declaration signatories. Accessed: 26.11.2020.
URL: *https://openaccess.mpg.de/319790/Signatories*

Maxwell, J. A. (2005). *Qualitative research design: An interactive approach*, volume 42. Sage publications, California, USA, 2nd edition. ISBN 978-1-4129-8119-4.

Maxwell, J. A. (2013). *Qualitative Research Design: An Interactive Approach*. SAGE Publications, Inc, Thousand Oaks, Calif, 3rd edition edition. ISBN 978-1-4129-8119-4.

Maxwell, J. A. and Loomis, D. M. (2003). Chapter 9 mixed methods design: An alternative approach. In *Handbook of mixed methods in social and behavioral research*, volume 1, pages 241–272. SAGE Publications, California, USA. ISBN 0-7619-2073-0.

Medjahed, B. and Bouguettaya, A. (2005). A multilevel composability model for semantic web services. *IEEE Transactions on Knowledge and Data Engineering*, 17(7):954–968. doi: 10.1109/TKDE.2005.101.

Meira, S. R. L., Buregio, V. A. A., Nascimento, L. M., Figueiredo, E., Neto, M., Encarnacao, B., and Garcia, V. C. (2011). The emerging web of social machines. In *2011 IEEE 35th Annual Computer Software and Applications Conference*, pages 26–27. doi: 10.1109/COMPSAC.2011.12.

Merriam, S. B. (1988). *Case study research in education: A qualitative approach*. Jossey-Bass. ISBN 978-1555423599.

Meschenmoser, P., Meuschke, N., Hotz, M., and Gipp, B. (2016). Scraping scientific web repositories: Challenges and solutions for automated content extraction. *D-Lib Magazine*, 22(9/10). doi: 10.1045/september2016-meschenmoser.

Meuser, M. and Nagel, U. (1991). ExpertInneninterviews—vielfach erprobt, wenig bedacht. In *Qualitativ-empirische sozialforschung*, pages 441–471. Springer. ISBN 978-3-322-93270-9.

Meuser, M. and Nagel, U. (2009). The expert interview and changes in knowledge production. In *Interviewing experts*, pages 17–42. Springer. ISBN 978-0-230-24427-6.

Miles, M. B., Huberman, A. M., and Saldaña, J. (2013). *Qualitative Data Analysis: A Methods Sourcebook*. SAGE Publications, Inc, Thousand Oaks, Califorinia, 3rd edition edition. ISBN 978-1-4522-5787-7.

Miller, P. (2000). Interoperability: What is it and why should i want it? Accessed: 03.04.2018.
URL: *http://www.ariadne.ac.uk/issue24/interoperability*

Millington, P. (2007). Counting on opendoar. Accessed: 15.05.2018.
URL: *https://slideplayer.com/slide/700137/*

Mingers, J. and Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, 246(1):1–19. doi: 10.1016/j.ejor.2015.04.002.

Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of arxiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, 58(13):2047–2054. doi: 10.1002/asi.20663.

Mongeon, P. and Paul-Hus, A. (2016). The journal coverage of web of science and scopus: a comparative analysis. *Scientometrics*, 106(1):213–228. doi: 10.1007/s11192-015-1765-5.

Morgan, C. (2008). Journal article version nomenclature: the niso/alpsp recommendations. *Learned Publishing*, 21(4):273–277. doi: doi.org/10.1087/095315108X356699.

Morse, J. M. (2003). Principles of mixed methods and multimethod research design. In *Handbook of Mixed Methods In Social and Behavioral Research*, pages 189–208. SAGE Publications.

Morse, J. M., Hupcey, J. E., Penrod, J., Spiers, J. A., Pooler, C., and Mitcham, C. (2002). Symposium conclusion - issues of validity: Behavioral concepts, their derivation and interpretation. *International Journal of Qualitative Methods*, 1(4):68–73. doi: 10.1177/160940690200100409.

Mulaik, S. A. (1985). Exploratory statistics and empiricism. *Philosophy of Science*, 52(3):410–430. doi: 10.1086/289258.

Müller, U., Severiens, T., Malitz, R., and Schirmbacher, P. (2009). Oa network: An integrative open access infrastructure for germany. *D-Lib Magazine*, 15(9):10. doi: 10.1045/september2009-mueller.

Murray-Rust, D. and Robertson, D. (2014). Lscitter: Building social machines by augmenting existing social networks with interaction models. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 875–880. ACM. doi: 10.1145/2567948.2578832.

Murray-Rust, D., Scekic, O., Truong, H.-L., Robertson, D., and Dustdar, S. (2014). A collaboration model for community-based software development with social machines. In *Proceedings of the 10th IEEE International Conference on*

Collaborative Computing: Networking, Applications and Worksharing. ICST. doi: 10.4108/icst.collaboratecom.2014.257245.

Murray-Rust, D., Scekic, O., Truong, H.-L., Robertson, D., and Dustdar, S. (2015a). A collaboration model for community-based Software Development with social machines. In *CollaborateCom 2014 - Proceedings of the 10th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pages 84–93. Institute of Electrical and Electronics Engineers Inc. ISBN 9781631900433. doi: 10.4108/icst.collaboratecom.2014.257245.

Murray-Rust, D., Tarte, S., Hartswood, M., and Green, O. (2015b). On wayfaring in social machines. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1143–1148. ACM. doi: 10.1145/2740908.2743971.

Myers, M. D. (2013). *Qualitative research in business and management*. Sage Publications, London, UK. ISBN 978-1-4129-2165-7.

Nascimento, L. M. d., Burégio, V. A., Garcia, V. C., and Meira, S. R. (2014). A new architecture description language for social machines. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 873–874. doi: doi.org/10.1145/2567948.2578831.

Ness, L. R. et al. (2015). Are we there yet? data saturation in qualitative research. *The Qualitative Report*, 20(9):1408–1416.

Nicholson, S. (2006). The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. *Information Processing & Management*, 42(3):785 – 804. ISSN 0306-4573. doi: 10.1016/j.ipm.2005.05.008.

Nicholson, D. and Dobreva, M. (2009). Beyond oais: towards a reliable and consistent digital preservation implementation framework. In *2009 16th International Conference on Digital Signal Processing*, pages 1–8. IEEE. doi: 10.1109/ICDSP.2009.5201126.

Nkiko, C., Bolu, C., and Michael-Onuoha, H. (2014). Managing a sustainable institutional repository: the covenant university experience. *Samaru journal of information studies*, 14(1-2):1–6.

Noorden, R. V. (2013). Open access: The true cost of science publishing. *Nature*, 495(7442):426–429. doi: 10.1038/495426a.

North, C. (2006). Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9. doi: 10.1109/MCG.2006.70.

OBrien, P., Arlitsch, K., Mixter, J., Wheeler, J., and Sterman, L. B. (2017). Ramp–the repository analytics and metrics portal: A prototype web service that accurately counts item downloads from institutional repositories. *Library Hi Tech*, 35(1):144–158. doi: 10.1108/LHT-11-2016-0122.

O'Hara, K. (2013). Social machine politics are here to stay. *IEEE Internet Computing*, 17(2):87–90. doi: 10.1109/MIC.2013.36.

Orlikowski, W. J. and Baroudi, J. J. (1991). Studying information technology in organizations: Research approaches and assumptions. *Information systems research*, 2(1):1–28. doi: 10.1287/isre.2.1.1.

Ottaviani, J. and Hank, C. (2009). Libraries Should Lead the Institutional Repository Initiative and Institutional Repositories : The Great Debate Institutional Repositories : The Great Debate. *Bulletin of the American Society for Information Science and Technology*, 35(4):17–21. doi: 10.1002/bult.2009.1720350408.

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.-J., Gundlach, J., Schirmbacher, P., and Dierolf, U. (2013). Making research data repositories visible: The re3data. org registry. *PloS one*, 8(11):e78080. doi: 10.1371/journal.pone.0078080.

Pardede, E., Rahayu, J. W., and Taniar, D. (2008). Xml data update management in xml-enabled database. *Journal of Computer and System Sciences*, 74(2):170–195. doi: 10.1016/j.jcss.2007.04.008.

Park, B.-K., Han, H., and Song, I.-Y. (2005). Xml-olap: a multidimensional analysis framework for xml warehouses. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 32–42. Springer. doi: 10.1007/11546849_4.

Patro, S. and Sahu, K. K. (2015). Normalization: A preprocessing stage. *International Advanced Research Journal in Science, Engineering and Technology*, 2(3). doi: 10.17148/IARJSET.2015.2305.

Patton, M. Q. (2002). *Qualitative Research & Evaluation Methods*. Sage Publications, New Delhi, India, 3rd edition. ISBN 0-7619-1971-6.

Pellack, L. (2018). Library Guides: Scopus: Comparisons. Accessed: 06.12.2018.
URL: *https://tinyurl.com/ybuhl6ap*

Pfadenhauer, M. (2009). At eye level: The expert interview - a talk between exper and quasi-expert. In *Interviewing experts*, pages 81–97. Springer. ISBN 978-0-230-24427-6.

Phillips, M. and Knoppers, B. M. (2019). Whose commons? data protection as a legal limit of open science. *The Journal of Law, Medicine & Ethics*, 47(1):106–111. doi: 10.1177/1073110519840489.

Pickton, M., Hitchcock, S., Coles, S., Morris, D., and Meece, S. (2010). Preserving repository content: practical steps for repository managers. In *The 5th International Conference on Open Repositories (OR2010)*. Madrid, Spain.

Pinfield, S. (2001). How do physicists use an e-print archive? implications for institutional e-print services. *D-Lib Magazine*, 7(12). doi: 10.1045/december2001-pinfield.

Pinfield, S. (2005a). A mandate to self archive? the role of open access institutional repositories. *Serials*, 18(1):30–34. doi: 10.1629/1830.

Pinfield, S. (2005b). Self archiving publications. In Gorman, G. and Rowland, F., editors, *International Yearbook of Library and Information Management 2004–2005: Scholarly Publishing in an Electronic Era*, pages 118–145. Facet Publishing, London, UK. ISBN 1-85604-536-6.

Pinfield, S. (2009). Journals and repositories: an evolving relationship? *Learned Publishing*, 22(3):165–175. doi: doi.org/10.1087/2009302.

Pinfield, S., Salter, J., Bath, P. A., Hubbard, B., Millington, P., Anders, J. H., and Hussain, A. (2014). Open-access repositories worldwide, 2005–2012: Past growth, current characteristics, and future possibilities. *Journal of the association for information science and technology*, 65(12):2404–2421. doi: doi.org/10.1002/asi.23131.

Pontika, N., Knoth, P., Cancellieri, M., and Pearce, S. (2015). Fostering open science to research using a taxonomy and an eLearning portal. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*, i-KNOW '15, pages 1–8. Association for Computing Machinery, New York, NY, USA. ISBN 978-1-4503-3721-2. doi: 10.1145/2809563.2809571.

Pontika, N., Knoth, P., Cancellieri, M., and Pearce, S. (2016). Developing infrastructure to support closer collaboration of aggregators with open repositories. *Liber Quarterly*, 25(4):172–188. doi: doi.org/10.18352/lq.10138.

Poynder, R. (2011). Open access by numbers. Accessed: 15.09.2018.
URL: *https://tinyurl.com/y4ltvna2*

Rabe, M. (2003). Revisiting'insiders' and'outsiders' as social researchers. *African Sociological Review/Revue Africaine de Sociologie*, 7(2):149–161. doi: 10.4314/asr.v7i2.23150.

Reese, T. (2009). Automated metadata harvesting: low-barrier marc record generation from oai-pmh repository stores using marcedit. *Library Resources & Technical Services*, 53(2):121–135. doi: 10.5860/lrts.53n2.121.

Rettberg, N. and Schmidt, B. (2012). Openaire-building a collaborative open access infrastructure for european researchers. *Liber Quarterly*, 22(3). doi: 10.18352/lq.8110.

Ribeiro, L., De Castro, P., and Mennielli, M. (2016). Eunis-eurocris joint survey on cris and ir: Final report. Technical report, EUNIS and euroCris.

RLG-OCLC (2002). Trusted digital repositories: Attributes and responsibilities. Technical report, RLG, California, USA.

ROARMAP (2010). The registry of open access repository mandates and policies (roarmap). Accessed: 27.10.2019.
URL: *https://jisc.ac.uk/rd/projects/monitoring-open-access-activity*

Robinson, M. and Horstmann, W. (2007). Driver-supporting institutional repositories in europe. In *Proceedings of the 11th International Conference on Electronic Publishing*, pages 445–446. Vienna, Austria. ISBN 978-3-85437-292-9.

Robson, C. and McCartan, K. (2016). *Real world research*. John Wiley & Sons, West Sussex, UK, 4 edition. ISBN 978-1-118-74523-6.

Rodden, T., Cheverst, K., Davies, K., and Dix, A. (1998). Exploiting context in hci design for mobile systems. In *First workshop on human computer interaction with mobile devices*.
URL: *https://tinyurl.com/y3mh3mgl*

Rodrigues, E., Shearer, K., et al. (2017). Next generation repositories: Behaviours and technical recommendations of the coar next generation repositories working group. Technical report, the Confederation of Open Access Repositories (COAR).

Ropohl, G. (1999). Philosophy of socio-technical systems. *Techné: Research in Philosophy and Technology*, 4(3):186–194. doi: 10.5840/techne19994311.

Rose, R. (2016). Defining analytics: A conceptual framework. *OR/MS Today*, 43(3).
URL: *https://tinyurl.com/ue7cuts*

Ross, J., Irani, L., Silberman, M., Zaldivar, A., and Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*, pages 2863–2872. ACM. doi: 10.1145/1753846.1753873.

Routio, P. (2007). Arteology, the science of products and professions. Viewed 2018-10-25.
URL: *http://www2.uiah.fi/projects/metodi/e00.htm*

Rush, J. E. (1996). *Scholarly publishing: The electronic frontier*. The MIT Press. ISBN 978-0262161572.

Russell, R. and Day, M. (2010). Institutional repository interaction with research users: a review of current practice. *New review of academic librarianship*, 16(S1):116–131. doi: 10.1080/13614533.2010.509996.

Sacha, D., Stoffel, A., Stoffel, F., Kwon, B. C., Ellis, G., and Keim, D. A. (2014). Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1604–1613. doi: 10.1109/tvcg.2014.2346481.

Saraiya, P., North, C., and Duca, K. (2005). An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):443–456. doi: 10.1109/TVCG.2005.53.

Satzger, B., Psaier, H., Schall, D., and Dustdar, S. (2013). Auction-based crowdsourcing supporting skill management. *Information Systems*, 38(4):547–560. doi: 10.1016/j.is.2012.09.003.

Schmidt, B., Bertino, A., Beucke, D., Brinken, H., Jahn, N., Matthias, L., Mimkes, J., Müller, K., Orth, A., and Bargheer, M. (2018). Open science support as a portfolio of services and projects: From awareness to engagement. *Publications*, 6(2):27. doi: 10.3390/publications6020027.

Scott, P. (2009). Institutional repositories: Content and culture in an open access environment. *Library Management*, 30(4/5):354–356. doi: 10.1108/01435120910958075.

Shadbolt, N., Brody, T., Carr, L., and Harnad, S. (2006). The open research web: A preview of the optimal and the inevitable. In *Open Access: Key Strategic, Technical and Economic Aspects*. Chandos Publishing (Oxford) Limited, Oxford, UK. ISBN 978-1-84334-203-8.

Shadbolt, N. R., Smith, D. A., Simperl, E., Van Kleek, M., Yang, Y., and Hall, W. (2013). Towards a classification framework for social machines. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, pages 905–912. Association for Computing Machinery, New York, NY, USA. ISBN 978-1-4503-2038-2. doi: 10.1145/2487788.2488078.

Shadbolt, N., O'Hara, K., De Roure, D., and Hall, W. (2019). *The Theory and Practice of Social Machines*. Lecture Notes in Social Networks. Springer. ISBN 978-3030108892.

Siguenza-Guzman, L., Saquicela, V., Avila-Ordóñez, E., Vandewalle, J., and Cattrysse, D. (2015). Literature review of data mining applications in academic libraries. *The Journal of Academic Librarianship*, 41(4):499 – 510. ISSN 0099-1333. doi: 10.1016/j.acalib.2015.06.007.

Smart, P. (2018a). Journal editors and data: the general data protection regulation (gdpr). *Journal: European Science Editing*, 44(3):50–51. doi: 10.20316/ESE.2018.44.18011.

Smart, P. (2018b). Knowledge machines. *The Knowledge Engineering Review*, 33(e11). **URL:** *https://philpapers.org/rec/SMAKM*

Smart, P., Simperl, E., and Shadbolt, N. (2014). A Taxonomic Framework for Social Machines. In Miorandi, D., Maltese, V., Rovatsos, M., Nijholt, A., and Stewart, J., editors, *Social Collective Intelligence: Combining the Powers of Humans and Machines to Build a Smarter Society*, chapter 3, pages 51–85. Springer International Publishing, Cham. ISBN 978-3-319-08681-1. doi: 10.1007/978-3-319-08681-1_3.

Smart, P., Madaan, A., and Hall, W. (2019). Where the smart things are: social machines and the internet of things. *Phenomenology and the Cognitive Sciences*, 18(3):551–575. doi: 10.1007/s11097-018-9583-x.

Suber, P. (2006). Open access in the united states. In *Open access: Key strategic, technical and economic aspects*, pages 149–160. Chandos Publishing. ISBN 978-1-84334-203-8.

Suber, P. (2007). Open access overview. https://tinyurl.com/b6e6h2f. Viewed 01.05.2018.

Suber, P. (2008). The open access mandate at harvard. *SPARC Open Access Newsletter*.
URL: *https://tinyurl.com/s7suf82*

Suber, P. (2009). Timeline of the open access movement. Accessed: 16.05.2019.
URL: *https://tinyurl.com/ttyyxm4*

Suber, P. (2012). *Open Access*. MIT Press Essential Knowledge, London, England. ISBN 978-0-262-51763-8.

Suleman, H. (2006). Parallelising harvesting. In Sugimoto, S., Hunter, J., Rauber, A., and Morishima, A., editors, *Digital Libraries: Achievements, Challenges and Opportunities*, pages 81–90. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-49377-8.

Swan, A. (2006). Overview of scholarly communication. In Jacobs, N., editor, *Open access: Key strategic, technical and economic aspects*, chapter 1, pages 3–12. Elsevier. ISBN 9781780632117.

Swan, A. (2010). The open access citation advantage: Studies and results to date. Accessed: 013.06.2018.
URL: *https://eprints.soton.ac.uk/268516/*

Swan, A. (2012). Institutional repositories–now and next. In *University libraries and digital learning environments*, pages 145–160. Ashgate Publishing. ISBN 978-1409486565.

Swan, A., Gargouri, Y., Hunt, M., and Harnad, S. (2015). Open access policy: Numbers, analysis, effectiveness. Technical report, Pasteur4OA Work Package 3 report: Open Access policies.
URL: *https://eprints.soton.ac.uk/375854/*

Tarte, S., Willcox, P., Glaser, H., and De Roure, D. (2015). Archetypal narratives in social machines: approaching sociality through prosopography. In *Proceedings of the ACM web science conference*, pages 24–34. ACM. doi: 10.1145/2786451.2786471.

Tennan, J. P., Crane, H., Crick, T., Davila, J., Enkhbayar, A., Havemann, J., Kramer, B., Martin, R., Masuzzo, P., Nobes, A., Rice, C., Rivera-López, B., Ross-Hellauer, T., Sattler, S., Thacker, P. D., and Vanholsbeeck, M. (2019). Ten hot topics around scholarly publishing. *Bibliosphere*, 7(2):3–25. doi: 10.20913/1815-3186-2019-3-3-25.

Thomas, J. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE. ISBN 9780769523231.
URL: *https://books.google.co.uk/books?id=DybZPAAACAAJ*

Tinati, R. and Carr, L. (2012). Understanding social machines. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 975–976. doi: 10.1109/SocialCom-PASSAT.2012.25.

Tinati, R., Wang, X., Tiropanis, T., and Hall, W. (2015). Building a real-time web observatory. *IEEE Internet Computing*, 19(6):36–45. doi: 10.1109/MIC.2015.94.

Tiropanis, T. (2012). A definition of the web observatory. Accessed: 11.12.2016.
URL: *http://www.webscience.org/a-definition-of-the-web-observatory*

Tiropanis, T., Hall, W., Shadbolt, N., De Roure, D., Contractor, N., and Hendler, J. (2013). The web science observatory. *IEEE Intelligent Systems*, 28(2):100–104. doi: 10.1109/MIS.2013.50.

Tiropanis, T., Hall, W., Hendler, J., and de Larrinaga, C. (2014a). The web observatory: a middle layer for broad data. *Big Data*, 2(3):129–133. doi: 10.1089/big.2014.0035.

Tiropanis, T., Rowland-Campbell, A., and Hall, W. (2014b). Government as a social machine in an ecosystem. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 903–904. ACM. doi: 10.1145/2567948.2578837.

Tiropanis, T., Wang, X., Tinati, R., and Hall, W. (2014c). Building a connected web observatory: architecture and challenges. In *2nd International Workshop on Building Web Observatories (B-WOW14), ACM Web Science Conference 2014*.
URL: *https://eprints.soton.ac.uk/366270/*

Torraco, R. J. (2005). Writing integrative literature reviews: Guidelines and examples. *Human resource development review*, 4(3):356–367. doi: 10.1177/1534484305278283.

Turner III, D. W. (2010). Qualitative interview design: A practical guide for novice investigators. *The qualitative report*, 15(3):754–760.

UKRI (2019). Ukri infrastructure roadmap progress report. Progress report, UK Research and Innovation, United Kingdom.

Unluer, S. (2012). Being an insider researcher while conducting case study research. *Qualitative Report*, 17(29):1–14.

Van de Sompel, H. and Nelson, M. L. (2015). Reminiscing about 15 years of interoperability efforts. *D-Lib Magazine*, 21(11-12). doi: 10.1045/november2015-vandesompel.

Van de Sompel, H. and Treloar, A. (2014). A perspective on archiving the scholarly web. In *Proceedings of the 11th International Conference on Digital Preservation*, pages 194–198. Melbourne. ISBN 978-0-642-27881-4.

Van de Sompel, H., Nelson, M. L., Lagoze, C., and Warner, S. (2004). Resource harvesting within the oai-pmh framework. *D-Lib Magazine*, 10(12). doi: 10.1045/december2004-vandesompel.

Vincent-Lamarre, P., Boivin, J., Gargouri, Y., Larivière, V., and Harnad, S. (2016). Estimating open access mandate effectiveness: The melibea score. *Journal of the Association for Information Science and Technology*. doi: 10.1002/asi.23601.

Walker, L. O., Avant, K. C., et al. (2005). *Strategies for theory construction in nursing*. Pearson/Prentice Hall Upper Saddle River, NJ, Boston, Massachusetts, USA, 4th edition. ISBN 978-0131191266.

Walsham, G. (1993). *Interpreting information systems in organizations*. John Wiley & Sons. ISBN 978-0471938149.

Walters, T. O. (2007). Reinventing the library how repositories are causing librarians to rethink their professional roles. *portal: Libraries and the Academy*, 7(2):213–225. doi: 10.1353/pla.2007.0023.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–33. doi: 10.1080/07421222.1996.11518099.

Wang, R. Y., Kon, H. B., and Madnick, S. E. (1993). Data quality requirements analysis and modeling. In *Proceedings of IEEE 9th International Conference on Data Engineering*, pages 670–677. doi: 10.1109/ICDE.1993.344012.

Wates, E. and Campbell, R. (2007). Author's version vs. publisher's version: an analysis of the copy-editing function. *Learned publishing*, 20(2):121–129. doi: doi.org/10.1087/174148507X185090.

Wegryn, G. (2014). Top 5 analytics predictions for 2015 from informs analytics section leader, informs podcast. Accessed: 12.4.2018.
URL: *https://tinyurl.com/y8p83s4c*

Weiss, R. S. (1995). *Learning from strangers: The art and method of qualitative interview studies*. Simon and Schuster. ISBN 978-0684823126.

Wickham, J. (2010). Repository management: an emerging profession in the information sector. In *Online Information 2010*.
URL: *http://eprints.nottingham.ac.uk/1511/*

Wijk, J. J. v. (2005). The value of visualization. In *VIS 05. IEEE Visualization, 2005.*, pages 79–86. IEEE. doi: 10.1109/VISUAL.2005.1532781.

Wilder, C. R. and Ozgur, C. O. (2015). Business analytics curriculum for undergraduate majors. *INFORMS Transactions on Education*, 15(2):180–187. doi: 10.1287/ited.2014.0134.

Williams, K., Wu, J., Choudhury, S. R., Khabsa, M., and Giles, C. L. (2014). Scholarly big data information extraction and integration in the citeseer χ digital library. In *Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on*, pages 68–73. IEEE. doi: 10.1109/ICDEW.2014.6818305.

Willinsky, J. (2003). The nine flavours of open access scholarly publishing. *Journal of Postgraduate Medicine*, 49(3):263–267.

Wolf, M. and Wicksteed, C. (1997). Date and time formats. Accessed: 01.05.2018.
URL: *https://www.w3.org/TR/NOTE-datetime*

Wong, P. C. and Thomas, J. (2004). Visual analytics. *IEEE Computer Graphics and Applications*, 24(5):20–21. doi: 10.1109/MCG.2004.39.

Wu, J., William, K., Chen, H.-H., Khabsa, M., Caragea, C., Tuarob, S., Ororbia, A., Jordan, D., Mitra, P., and Giles, C. L. (2015). Citeseerx: Ai in a digital library search engine. *AI Magazine*, 36(3):35–49. doi: 10.1609/aimag.v36i3.2601.

Xia, F., Wang, W., Bekele, T. M., and Liu, H. (2017a). Big scholarly data: A survey. *IEEE Transactions on Big Data*, 3(1):18–35. doi: 10.1109/TBDATA.2016.2641460.

Xia, F., Wang, W., Bekele, T. M., and Liu, H. (2017b). Big scholarly data: A survey. *IEEE Transactions on Big Data*, 3(1):18–35. doi: 10.1109/TBDATA.2016.2641460.

Yang, J., Redi, J., Demartini, G., and Bozzon, A. (2016). Modeling task complexity in crowdsourcing. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, pages 249–258.

Yi, J. S., Kang, Y.-a., Stasko, J. T., and Jacko, J. A. (2008). Understanding and characterizing insights: how do people gain insights using information visualization? In *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, page 4. ACM. doi: 10.1145/1377966.1377971.

Yin, R. (2009). *Case Study Research and Applications: Design and Methods*. SAGE Publications, 4 edition. ISBN 978-1-4129-6099-1.

Yin, R. K. (2011). *Applications of case study research*. Sage, 3 edition. ISBN 978-1412989169.

Yu, S. and Woodard, C. J. (2009). Innovation in the programmable web: Characterizing the mashup ecosystem. In Feuerlicht, G. and Lamersdorf, W., editors, *Service-Oriented Computing – ICSOC 2008 Workshops*, pages 136–147. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-642-01247-1. doi: doi.org/10.1007/978-3-642-01247-1_13.

Zhai, Y. and Liu, B. (2006). Structured data extraction from the web based on partial tree alignment. *IEEE Transactions on Knowledge and Data Engineering*, 18(12):1614–1628. doi: 10.1109/TKDE.2006.197.

Zheng, H., Li, D., and Hou, W. (2011). Task design, motivation, and participation in crowdsourcing contests. *International Journal of Electronic Commerce*, 15(4):57–88. doi: 10.2753/JEC1086-4415150402.

Zimmerman, E. H. (2002). Cris-cross: Research information systems at a crossroads. In *6th International Conference on Current Research Information Systems*. euroCRIS, Kassel, Germany.
URL: *http://hdl.handle.net/11366/129*

Zuccala, A., Oppenheim, C., and Dhiensa, R. (2008). Managing and evaluating digital repositories. *Information Research: An International Electronic Journal*, 13(1).
URL: *http://informationr.net/ir/13-1/paper333*

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., and Alibaks, R. S. (2012). Socio-technical impediments of open data. *Electronic Journal of e-Government*, 10(2):156–172.