Mind the Gap! Moving From Aspiration to Experience in UK Institutional Research Data Management

Leslie Carr
University of Southampton, UK, lac@ecs.soton.ac.uk

BACKGROUND

In the wake of the 10th anniversary of the Open Archiving Initiative there is a natural opportunity to examine the normal practice of research staff in institutional contexts and seriously consider how to their research data can be practically managed in institutional repositories – to elevate research data to be a first-class citizen in the world of open scholarly communication and to mainstream some aspects of e-research. Such a goal requires far more than technical capability, but encompasses significant change for all stakeholders. The aim of the Institutional Data Management Blueprint (IDMB) project [1], funded by JISC in the UK from 2009 to 2011, has been to create a *practical and attainable* institutional framework for managing research data that facilitates *ambitious* e-research practice. The goal is to produce a framework for managing the research data of a whole institution informed by an analysis of current data management requirements for a representative group of disciplines and to pilot an implementation plan for an institution-wide data model, that is integrated into existing research workflows and that extend the potential of existing data storage systems, including those linked to discipline and national shared service initiatives.

Defining the responsibilities of data management from inception to preservation is now clearly recognised as a complex process shared between individual researchers and research groups, institutions, funders and national agencies [2], driven by many agendas, including groups of users, different funding agencies and programmes, politics, and technology trendsetters. Within this group of stakeholders the institution can act as a centre for cohesion, curation and cooperation – an agent that can assume responsibility for its own research data at some, or maybe all, of its lifetime. A candidate tool to support this responsibility is the institutional repository – an information storage and management tool conjoined with extensive social support and advice structures from the library. In order to acknowledge and manage their data management responsibilities, IDMB provides an overall framework within which to plan and develop institutional data management strategy. Many of the landscape studies so far have been highly detailed analytical descriptors of theoretical models, applicable only to specific disciplines, which institutions can find difficult to implement, and which can be too complex to win engagement from researchers.

This rest of this paper describes the main practical developments being made to an institutional repository platform as a result of the IDMB data management survey and audit [3].

REPOSITORY

The University of Southampton Institutional Repository is based on the EPrints platform (v. 3.2), configured for some rudimentary data support: 'dataset' records are differentiated from publication (research output) records with a separate workflow, some extra metadata (*e.g.* ethical clearance) and common kinds of data files explicitly recognized and tagged (spreadsheets, database files, math and statistics packages). The net effect is that research data is discoverable, but not easily interpretable or reusable. A table of data points may be provided as a spreadsheet, a database or a PDF, but guidance as to the interpretation of those figures is not easy to come by. Nor is it easy to understand the relationship between multiple data files (components of complex data objects.)

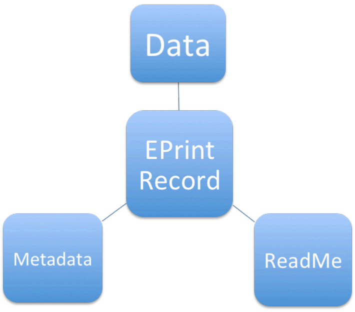


Figure 1: Document Contents

As a result of the IDMB project, the following changes have been made to the repository platform to better support data capture:

The content metadata property of documents[1], normally used to distinguish between published, submitted, draft and related material, has been extended to also identify *data*, *software*, *metadata* and *explanatory (aka README)* material.

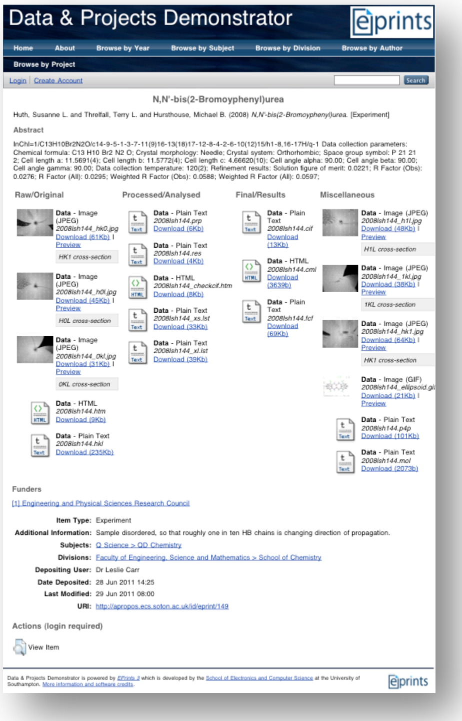
An eprint record of type *dataset* might contain three documents:

1. a database document (content=data) that provides the actual data of the dataset,
2. a word processor document (content=README) that describes for a human reader the details of the interpretation of the spreadsheet's rows and columns,
3. a spreadsheet document (content = metadata) that contains descriptive or contextual information about the dataset (*e.g.* machine settings and experimental conditions under which the data was gathered) using attribute/value pairs from a discipline-specific metadata schema.

[1] an eprint record can contain zero or more documents; each document contains one or more files. Both eprints and documents have extensive semantic metadata whereas files only have storage-related metadata. A common example of multi-file documents are HTML pages; but most EPrints repository documents are (of course) single-file PDFs or Office documents.

Where the explanatory information does not require rich display media, README information can be provided in a bespoke metadata field for data documents, simplifying the deposit process.

To support experimental activities consisting of various stages of investigation and analysis an extra metadata property is provided in the document ingest workflow. Each data file can be attributed to a specific *stage* – the default set of stages is defined to be applicable across many disciplines: data collection, data analysis, and results. A more complex experimental environment with standardized practices (*e.g.* crystallography) has the opportunity to provide an OAI-ORE RDF file (content=README) that explains the specific relationship between all the data documents and the formal model of the actual experimental procedure [4]. Figure 2 shows a crystallography experiment in the project data repository: the repository metadata groups the data files into the three default stages for human readers, whereas the attached ORE RDF document elaborates the complete complexity of the experimental process for e-research agents.



N,N'-bis(2-Bromoyphenyl)urea

Huth, Susanne L. and Threlfall, Terry L. and Hursthouse, Michael B. (2008) *N,N'-bis(2-Bromoyphenyl)urea.* [Experiment]

Abstract

InChl=1/C13H10Br2N2O/c14-9-5-1-3-7-11(9)16-13(18)17-12-8-4-2-6-10(12)15/h1-8,16-17H/q-1 Data collection parameters: Chemical formula: C13 H10 Br2 N2 O; Crystal morphology: Needle; Crystal system: Orthorhombic; Space group symbol: P 21 21 2; Cell length a: 11.5691(4); Cell length b: 11.5772(4); Cell length c: 4.66620(10); Cell angle alpha: 90.00; Cell angle beta: 90.00; Cell angle gamma: 90.00; Data collection temperature: 120(2); Refinement results: Solution figure of merit: 0.0221; R Factor (Obs): 0.0276; R Factor (All): 0.0295; Weighted R Factor (Obs): 0.0588; Weighted R Factor (All): 0.0597;

Raw/Original

Data - Image (JPEG) *2008lsh144_hk0.jpg* Download (61Kb) I Preview
HK1 cross-section

Data - Image (JPEG) *2008lsh144_h0l.jpg* Download (45Kb) I Preview
H0L cross-section

Data - Image (JPEG) *2008lsh144_0kl.jpg* Download (31Kb) I Preview
0KL cross-section

Data - HTML *2008lsh144.htm* Download (9Kb)

Data - Plain Text *2008lsh144.hkl* Download (235Kb)

Processed/Analysed

Data - Plain Text *2008lsh144.prp* Download (6Kb)

Data - Plain Text *2008lsh144.res* Download (4Kb)

Data - HTML *2008lsh144_checkcif.htm* Download (8Kb)

Data - Plain Text *2008lsh144_xs.lst* Download (33Kb)

Data - Plain Text *2008lsh144_xl.lst* Download (39Kb)

Final/Results

Data - Plain Text *2008lsh144.cif* Download (13Kb)

Data - HTML *2008lsh144.cml* Download (3639b)

Data - Plain Text *2008lsh144.fcf* Download (69Kb)

Miscellaneous

Data - Image (JPEG) *2008lsh144_h1l.jpg* Download (48Kb) I Preview
H1L cross-section

Data - Image (JPEG) *2008lsh144_1kl.jpg* Download (38Kb) I Preview
1KL cross-section

Data - Image (JPEG) *2008lsh144_hk1.jpg* Download (64Kb) I Preview
HK1 cross-section

Data - Image (GIF) *2008lsh144_ellipsoid.gi* Download (21Kb) I Preview

Data - Plain Text *2008lsh144.p4p* Download (101Kb)

Data - Plain Text *2008lsh144.mol* Download (2073b)

Funders

[1] Engineering and Physical Sciences Research Council

Item Type: Experiment
Additional Information: Sample disordered, so that roughly one in ten HB chains is changing direction of propagation.
Subjects: Q Science > QD Chemistry
Divisions: Faculty of Engineering, Science and Mathematics > School of Chemistry
Depositing User: Dr Leslie Carr
Date Deposited: 28 Jun 2011 14:25
Last Modified: 29 Jun 2011 08:00
URI: http://apropos.ecs.soton.ac.uk/id/eprint/149

Actions (login required)

View Item

Data & Projects Demonstrator is powered by *EPrints 3* which is developed by the School of Electronics and Computer Science at the University of Southampton. More information and software credits.

Figure 2: Crystallography Experiment showing stages of simplified workflow & explanatory OAI-ORE readme file

This tradeoff between expressive power for adept e-research disciplines and simplicity of use for stakeholders with more conservative practice must be a key concern for an Institutional Repository.

CONCLUSIONS AND FUTURE WORK

Experience shows that some simple extensions to the EPrints repository platform make it possible to construct useful repository records in a way that mimics the functionality of established information systems (subject repositories, databases and web services) and that enables readers to interpret the contents and structure of the experimental data, but that it is not always efficient to do so.

Disciplines such as crystallography that have complex multi-stage experimental procedures with dozens of data files per experiment need automatic ingest procedures specific to their discipline. Such procedures will take advantage of the new SWORD 2.0 protocol [5]. It is not just complexity that provides a problem – bulk is equally problematic. Experimental equipment (such as the Nanofabrication department's Helium Ion microscope) that dumps set after set of image data into a drop folder on its controlling PC would benefit from a simple but lab-specific data ingester that pulls together the contextual metadata from experimental schedules and project databases with the 'latest' set of image dumps as a rich repository deposit, rather than a manual transfer to a USB stick.

We also aim to add Open Provenance (OPM) metadata to the repository to further facilitate the description of experimental workflow as well as describe the repository administration processes to support curation and preservation.

REFERENCES

1. Takeda, K., et al., *Data Management for All - The Institutional Data Management Blueprint project, in Proceedings of 6th International Digital Curation Conference . 2010.* Available from: http://www.southamptondata.org/uploads/7/3/0/0/730051/6th_international_digital_curation_conference__idmb_final_paper_revised.pdf, accessed 29 Jun 2011.
2. Lyon, L. Dealing with Data: Roles, Rights, Responsibilities and Relationships – UKOLN Consultancy Report. 2007
3. University of Southampton, IDMB Initial Findings Report, http://www.southamptondata.org/1/post/2010/12/idmb-initial-findings-report.html
4. Lagoze, Carl (2009) The oreChem Project: Integrating Chemistry Scholarship with the Semantic Web. In: Proceedings of the WebSci'09: Society On-Line, 18-20 March 2009, Athens, Greece.
5. Tarrant, D., Carr, L., Wade, A. and Warner, S. (2010) Interactive Multi-Submission Deposit Workflows for Desktop Applications. In: Open Repositories 2010, July 2010, Madrid, Spain.

ABOUT THE AUTHOR

Dr Leslie Carr is a senior lecturer in the Web and Internet Science research group at the University of Southampton, UK, and a director of the Web Science Doctoral Training Centre. Dr Carr is the director of EPrints (eprints.org), the first institutional repository platform established in 1999, which supports over 320 public repositories supporting open access, open data and open educational resources agendas. Dr Carr is also the Director of EPrints Services, a spinout of the University of Southampton that commercially exploits the open source EPrints software, providing repository services and training to the international research industry. In conjunction with the EPrints team at Southampton, Dr Carr has led over twenty funded digital library and repository projects including IDMB and DepositMO.