6th International Digital Curation Conference

December 2010

Data Management for All - The Institutional Data Management Blueprint project

Kenji Takeda[1], Mark Brown, Simon Coles, Les Carr, Graeme Earl, Jeremy Frey, Peter Hancock, Wendy White, Fiona Nichols, Michael Whitton, Harry Gibbs, Christine Fowler, Pam Wake, Steve Patterson

University of Southampton

December 2010

Abstract

In the 10th anniversary year of the Open Archiving Initiative it is necessary to elevate research data to be a first-class citizen in the world of open scholarly communication. Such a profound goal requires far more than technical capability, but encompasses significant change for all stakeholders. Data curation and data management is often seen as an additional task for researchers. In this paper we describe how we are attempting to make it a seamless part of a researcher's daily workflow across a wide range of disciplines as a cornerstone of research practice. This paper describes the Institutional Data Management Blueprint (IDMB) project, which aims to create a practical and attainable institutional framework for managing research data throughout its lifecycle that facilitates ambitious national and international e-research practice. The objective is to produce a framework for managing research data across the whole lifecycle that encompasses a whole institution (exemplified by the University of Southampton) and based on an analysis of current data management requirements for a representative group of disciplines with a range of different data.

Note: This paper is a summary of the IDMB Initial Findings report, which is available at www.southamptondata.org

[1] Corresponding author: ktakeda@soton.ac.uk

Introduction

There has been a great deal of work contributed to defining and scoping aspects of the research data lifecycle, a number of which have sought to engage directly with researchers, which is recognised as increasingly important. Defining the responsibilities for managing data 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]. This is driven by many agendas, including groups of users, different funding agencies and programmes, politics, and technology trendsetters. A constant factor is the institution - a centre for cohesion, curation and cooperation - which is responsible for its own research data at some, or maybe all, of its lifetime, within a fragmented and volatile world. In order to acknowledge and manage these responsibilities, institutions require an overall framework within which to plan and develop their data management strategy. Many of the landscape studies so far have been highly detailed analytical descriptors of theoretical models, with some testing of assumptions, which institutions can find difficult to implement, and which can be too complex to win engagement from researchers. The management of data requires a multifunctional team approach which can bring together the knowledge and expertise of both researchers and professionals within an institutional policy and technical framework.

This paper describes the Institutional Data Management Blueprint (IDMB) project, which aims to create a practical and attainable institutional framework for managing research data that facilitates ambitious national and international e-research practice.

Aims and Objectives

The objective is to produce a framework for managing research data across the whole lifecycle that encompasses a whole institution (exemplified by the University of Southampton) and based on an analysis of current data management requirements for a representative group of disciplines with a range of different data. Building on the developed policy and service-oriented computing framework, the project has scoped and is evaluating a pilot implementation plan for an institution-wide data model, which can be integrated into existing research workflows and extend the potential of existing data storage systems, including those linked to discipline and national shared service initiatives. The project builds upon a decade of previous open access repository initiatives at Southampton to create a coherent set of next actions for an institutional, cross-discipline 10-year roadmap, which will be flexible in accommodating future moves to shared services, and provide a seamless transition of data management to knowledge transfer, from the individual to the community and from the desktop to institutional, national and international repositories. The outcomes from this project, which draws together technical, organisational and professional expertise from across the institution, will be widely disseminated within the sector as a form of HEI Data Management "Business Plan How-To".

[2] http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/publications.html#2007-06-19

The final outcomes from the project will include the following:

- Pathfinder for institutional data management strategy for the next decade;
- Data management institutional blueprint based on an analysis of data management requirements and current best practice;
- Service-oriented, extensible enterprise architecture model for data management;
- 10-year business model roadmap;
- Best practice gap analysis report;
- Pilot implementation for infrastructure, human and technological;
- Workshops, training, website and reports for dissemination of best practice.

The realisable benefits to the institution and the wider community are the following:

- Coherent data management strategy for a single institution;
- Change management strategy for open access of data;
- Development of cross-professional skills base for managing research data, including graduate student training;
- Preservation and curation of research data at an institutional level;
- Advocacy and best practice guidance for research workflows and data across disciplines;
- Cost-benefit analysis of implementing the framework;
- Detailed business model blueprint for others, to accelerate early adoption.

Data Management Audit

A key part of the IDMB project is to engage with users to ascertain current data management practice, support and constraints. In order to do this we have extended the data management audit carried out for the Southampton School of Social Sciences as part of the DataShare project[3]. This was carried out alongside face-to-face workshops with the research community.

[3] Gibbs, H., (2009), Southampton Data Survey: Our Experience and Lessons Learned, DISC-UK DataShare project

The audit involved a four-pronged approach to gather quantitative and qualitative data at the individual, School and University levels:

- **Online questionnaire.** This was used to provide quantitative information and some base level qualitative information across a spectrum of areas including current practice, policy, and governance. These were targeted at individuals in the Schools of Electronics & Computer Science, Engineering Sciences and Humanities.
- **Interviews.** Follow-up interviews with willing questionnaire respondents were used to obtain more details from individuals to provide more qualitative data. This allowed us to drill-down into specific area that participants were particularly interested/concerned with.
- **AIDA (Assessing Institutional Data Assets) tool[4].** The AIDA self-assessment tool is designed to provide benchmarking data on the level of data management capability available. Here it has been applied at the School level.
- **Crowdsourcing.** In order to obtain additional feedback, and experimental crowd-sourcing approach has been piloted. This is using the project website (www.southamptondata.org) and uses an *ideas box* approach.

In this section we describe the key findings from the questionnaire, interviews and AIDA benchmarking. Notably:

- Guidance and advice on research data management was limited;
- Knowledge of available capability and resources was limited;
- Researchers resorted to their own best efforts in many cases, e.g. USB hard drives;
- Data requirements are growing, almost half of respondents stored more than 100GB of research data;
- Most users had experienced problems due to lack of storage;
- Longevity of storage is considerable, mean 5 years, many researchers express preference for keeping research data 'forever';
- Backup practices were inconsistent, with users wanting better support for this;
- Researchers need help on how to organise their research data;
- Many researchers share, or would like to share, their data;
- Many researchers use other people's data, particularly within their own group;
- There is considerable scope for improvement in the provision of resources and capability.

[4] http://aida.jiscinvolve.org/wp/

A modified version of the AIDA toolkit was used to perform benchmarking of the current status of research data management in the three Schools surveyed. While some concerns over the validity of the process for completing the AIDA survey were expressed, it has proven useful as a basic check. In addition to the findings above from the questionnaires and interviews, the following could be inferred from the AIDA process:

- Capabilities across different Schools varies, with pockets of best practice throughout;
- Schools research practice is embedded and unified;
- Most of the data management capability tends to be localized;
- Formalization of data management policies and procedures would be beneficial;
- Technological capability needs to be more uniformly supported at the institutional level;
- Resources are generally limited.

Data Management Framework

In this section we highlight the current data management framework at the University of Southampton. It is clear that there is a robust policy framework at the University of Southampton that encompasses data management, ownership, IPR and freedom of information. The current issue is that this information is scattered in a way that it is difficult for researchers to access in a coherent way. It therefore appears that guidance is disjointed, and that the policy framework is not coherent. Similarly, guidance on data management is not clearly signposted, points of contact not clearly identified, and areas of responsibility between professional services not readily apparent.

Data management infrastructure and services are being consolidated within iSolutions, although Schools still house local capability of significant capacity. There is a plethora of different data solutions, coupled with a general lack of capacity. This has stimulated researchers to find their own solutions. While some of these are being migrated to central systems, the cost of doing this across the board would be significant.

The provision of backup services that are affordable and easy-to-use is not readily apparent to researchers. Also, researchers make the distinction between reproducible and non-reproducible data, which they are willing to manage cost/risk against in terms of paying for services. This is not readily supported at an institutional level.

It is apparent that while central systems can provide better support, there will always be more specialised requirements. Therefore the future strategy must combine commonality for consistent and affordable solutions, with flexibility to meet researchers varying needs.

The data lifecycle, particularly for curation and preservation, is not clearly handled by the institution, both technically or organisationally.

There is a lack of formal training around data management, and limited self-help and guidance for researchers. In some areas there is exemplary best practice, and it is important

that this is shared and promoted across the university for the benefit of all. This is important as data management plans become more prevalent. This will help to ensure that researchers are as productive as possible, in order to meet the university's ambitious strategic goals.

In terms of metadata management, an institution-wide framework has been proposed around a three-layer metadata model. The use of Dublin Core, and development of qualified Dublin Core, is suggested as a way of standardising use of metadata while providing extensibility within disciplines. Using a common framework has advantages in terms of training and support across the institution, development and use of tools, and embedding common data management practice within the researcher's daily lives.

Metadata framework

The aim of the IDMB project is to provide some guidance and pilots for better data management across the University of Southampton.Clearly, the use of metadata is necessary to add meaning and context to research data. In order to try and create a metadata framework that is applicable across domains, we look at what the end user would want to achieve. As an example, we use the analogy of archiving a retiring professor's office in a day – there is not enough time to record/classify each object. Instead papers, etc. of a similar nature are placed into folders, and then into a box, with a label describing the contents in general terms. In future researchers can access material if they identify a box file. Once opened they can identify folders, and then, if of sufficient interest, they can then find the relevant papers to investigate in detail. We break this down into three levels of findability:

- Core metadata *(box file)*. In order to find author, publisher, discipline, date;
- Discipline metadata *(folder)*. To find the right sub-domain, project, funder, technique;
- Project metadata *(paper)*. To find detailed dataset and its context.

This three layer metadata approach is illustrated in Figure 1 with discipline examples for aeronautics and archaeology shown in Figure 2. If used appropriately we believe that this metadata model is able to satisfy the requirements of the user scenarios identified previously. This model provides flexibility for the creator, while trying to include applicability to the end user. The difficulty occurs when these people have different roles, such as a researcher as the creator, and an archivist as the end user.

One of the more developed standards for metadata is Dublin Core, which is a set of text-based elements that can be used to describe resources such as books, media, and data. It has been developed since 1995 and now defined as ISO standard 15836 and NISO Standard Z39.85-2007. The continued development of this is now through the Dublin Core Metadata Initiative (DCMI[5]). Simple Dublin Core comprises of fifteen basic elements, with Qualified Dublin Core including three additional elements to cover: audience; provenance and RightsHolder. Dublin Core is typically implemented in XML, and as such has been

[5] http://dublincore.org/

popularised for open access through the Open Archives Initiative for Metadata Harvesting (OAI-PMH)[6].

Here we propose that Dublin Core is an appropriate standard to be used for an institution-wide metadata framework to provide the first-level of metadata. This is the approach already used by the National Crystallography Centre at Southampton through eCrystals, and is supported by EPrints.

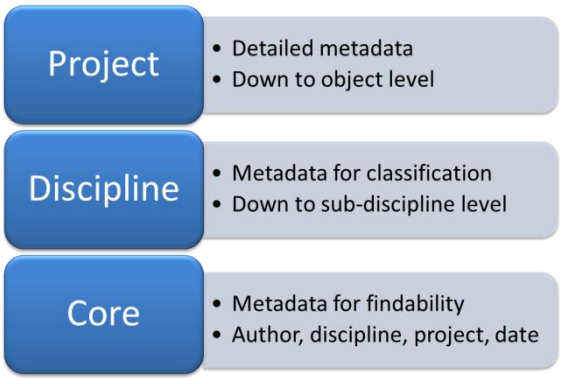

Figure 1. Three-layer metadata model

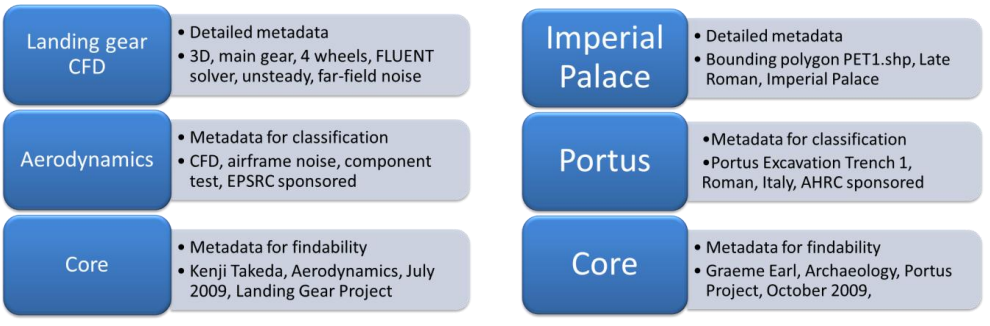

Figure 2. Three-layer metadata model examples for aeronautics (left) and archaeology (right).

Key challenges that face the uptake of better metadata management include:

- How to encourage people to tag their data;
- Metadata schemas that are not onerous;
- Usable tools for metadata assignment and import;
- Provenance tracking;
- Automating metadata assignment.

These will be all developed and addressed as part of the pilot studies in archaeology and the nano-fabrication centre.

[6] http://www.openarchives.org/OAI/openarchivesprotocol.html

Pilot Implementations

The initial phases of the IDMB project have highlighted many of the data management challenges faced by the institution. We have identified ways forwards, and will be embarking on three pilot implementations to see how more coherent, integrated, and intuitive data management solutions can be developed and deployed. These are in the areas of archaeology, the Nano-Fabrication Centre, and meta-search across federated repositories.

Archaeology

Archaeology researchers handle many different types of data, and cover a wide spectrum of requirements for data management that are applicable in other disciplines. In this pilot we will be exploring the use of Microsoft SharePoint 2010 as a virtual research environment, supporting researchers' data management needs.

User Scenario 1: Working with Geophysical Survey Data

David has a good understanding of technology. He has been doing geophysical surveys for many years and has developed his own best practice in working with the project data produced. Some of his work is commercial and some research has ethical constraints, therefore IP and other restrictions need to be carefully dealt with.

During his archaeological fieldwork he keeps all the survey files produced on his laptop. The software packages he is using require specific file structures on his C-drive. This has caused him some problems, as at the Archaeological Computing Lab, where there are more powerful computers to process data, he does not have sufficient user rights to keep files on the C-drive. David is very busy trying to finish his analysis and writing reports so there is not much time left for proper archiving after the project fieldwork is completed. He deposits all the project files on the iSolutions Archaeology shared project folder, which is backed up every night. He works very closely with colleagues in Italy. As he is responsible for the project he would like to have access to their data and vice versa, but he has not found a good solution for that. David is generally happy about the way of working with his project data, even though sometimes he dreams about an assistant who helps to keep his digital and paper archives better organised.

User Scenario 2: Working with Computer Graphics

Tim knows a lot about 3D graphics and building survey. After receiving his Master's degree he worked as a research assistant. During that time he was involved in several survey and 3D computer reconstruction projects. He manages his data by capturing surveys either on the instrument or direct to a laptop. These are then backed up to a memory stick and frequently emailed home whilst on fieldwork. Gathering the data in two different ways can lead to confusion. Also each CAD file produced frequently contains the previous days' surveys in addition to that produced on the current day. The 3D reconstruction work builds on the surveys and produced a wide range of architectural and landscape simulations. Whilst there are a range of proposed methods for documenting the processes and files involved there is no consensus

on the best approach. Currently Tim uses a hand written survey notebook and a typed 'diary' during his reconstruction work as the main documentation.

User Scenario 3: PhD Student Working from Home

Alison is a PhD student who mainly works at home. She finished her MSc of Archaeological Computing course several years ago. She also did her undergraduate degree in Southampton, so she is familiar with the infrastructure that the University provides. She has her data very well organised: she keeps paper copies of readings and notes, and also backs up all her files regularly on her external hard-drive. When she comes to work in university she brings her laptop with her, or sometimes she copies files to her G-Mail account as an attachment. Alison does not use iSolutions disk space as she finds it difficult and unreliable to access from home. Also, a few times when she tried this method all of her GIS data files somehow became corrupted. Alison makes use of online link and bibliographic management tools such as Zotero, Mendeley and Delicious. However, she feels that many of her colleagues would benefit from these but do not make use of them.

User Scenario 4: Senior Lecturer

Katie is senior lecturer. She works on an AHRC funded research field project. The AHRC requires deposit of a project archive with the Archaeology Data Service. As a consequence the data produced during the project must be consistent with deposit requirements, or be sufficiently documented and organised to enable the production of appropriate data and metadata with minimal additional investment. Preparation of data for deposit is rarely if ever built in as a major work package into Katie's work. As Katie wants to continue to develop her data beyond the lifecycle of her funded project she wants a means to expose her ongoing work in a way that makes it accessible and useful to others. As a consequence she requires that her data must as far as possible be exposed as RDF. Katie has three main needs for the documentation and management of her archaeological project data:

Image metadata:

On Katie's large field project image metadata is stored in EXIF and IPTC data and in an external metadata catalogue which enables CSV export. EXIF and IPTC data is attributed either direct via the camera (e.g. GPS spatial data, timestamp, creator, camera number) or via software such as AdobeBridge or download/ upload processing scripts. An automated tagging process or series of processes are run and checked prior to manual tagging. These data are managed centrally at the University of Southampton via a SharePoint server, using the Sharepoint Media Asset Library. At the end of each season data are uploaded to SharePoint. At the beginning of each season the data are downloaded from SharePoint and the data are locked on the server.

Temporal metadata:

Temporal data are gathered for many items on Katie's AHRC field project. A classic example is provided by Amphora data. Amphora sherds are recorded in the site office in Italy and entered into a database. The database will record the type of amphora. This type is in turn associated with various kinds of temporal information. The data and assets associated with these data should be attributed with appropriate temporal metadata and we will need to perform Allen operator based probabilistic reasoning ideally.

Spatial metadata:

Spatial data are also common on the project. Geographic Information Systems data include spatial information. The Sharepoint server must ingest these data and enable their display. Sharepoint is used to manage appropriate hierarchical spatial metadata. Note: the ArcGIS plugin can be used to attribute data with a location using a map as with Flick geocoding. Need to be able to see map data in any Sharepoint page.

User Scenario 5: Retired Professor

Kris is an emeritus professor at the department. He retired a few years ago but he still filled with energy and ideas. Most of his life he has been using pen and paper and occasionally a typewriter, so he has a very big archive in his office. He admires technology but he remains a novice.

He use computer to write articles and to prepare his presentations. Kris has never lost any of his files even though he keeps them all in the My Documents folder with no further folder structure. He is very organised with his paper records, with them all nicely filed in his drawers and on his bookshelves. He can find a note on a paper record from twenty years ago but would find it nearly impossible to find a similar digital note.

In addition to Kris's paper notes he believes that he has the only copies of a number of vital paper documents. For example, the archives from a number of excavations remain in his filing cabinets. Similarly he knows of some physical archives from his and colleagues' excavations that are in stores in the department. Some of these are not organised and are poorly labelled.

One thing Kris desperately needs is an easy way of sharing the articles and presentations he has with other people. He has heard of ePrints but not used it yet. He also wants to make some simple web pages to accompany his existing publications so that he does not have to conventionally publish many hundreds of plans and photographs. He doesn't have any research budget to pay for this so ideally needs a system that he can learn to use himself.

Nano-Fabrication Centre

The second pilot is for the Nanofabrication Centre (www.southampton-nanofab.com), the newly established state-of-the-art facility for nanofabrication and characterisation, run by the Nano Research Group from Electronics and Computer Science. An experimental data management repository will be established based on procedures related to two new pieces of equipment: the ASM Epsilon Epitaxy System and the Orion helium ion microscope. Currently, although data from each experiment is stored digitally by the machines, records of the experimental settings used and the outputs obtained from both these systems are maintained in student logbooks; exploration of the parameter space is achieved by coordinated sharing of paper records. Initial discussions have shown that an eCrystals-style repository storing raw data, sufficient metadata to recreate the experimental conditions, plus data from the intermediate stages of analysis will have the capability to yield a significantly positive effect on the laboratory procedure.

User Scenario 1: Helium ion microscope single inspection

John is a researcher a new silicon device in the Southampton Nanofabrication Centre. All users of the facility use the computer-based Clean Room Management System (CRMS) to plan their experiments. They can choose from a *recipe book* and modify parameters to fit their task. Alternatively they can start a new process from scratch, entering all of the relevant machine and process parameters starting fresh. John chooses a standard process to being manufacture of his device.

John creates his device and then must inspect it using the Orion helium ion microscope. This again uses the CRMS, but here he revises the entries to what the process actually ran –i.e. actual parameters, rather than requested ones. The microscope produces a series of images, which are then transferred to an EPrints data system. EPrints requests the metadata from the CRMS so that it is stored alongside with microscope data. Once the experiment is finished, John can return to his office and access his data files over the network via EPrints. John can now manage his microscope data in an organised way without having to worry about storage, backup or archive, as this is now taken care of centrally.

Meta-Search

The third pilot will develop a proof-of-concept demonstrator that enables cross-disciplinary data linking and researcher expertise matching. This is to enable transformative inter-disciplinary science that is currently difficult to achieve due to the discipline-specific silo nature of open data repositories. The user community targeted is the Southampton Nano-Forum, a University Strategic Research Group, aligned to the national EPSRC theme, that comprises researchers across electronics and computer science, chemistry, physics, engineering sciences, and mathematics. It is based on the twin observations that (a) related disciplines tend to have similar working practices and academic values and (b) it is quicker to establish e-research and repository services at a departmental or subject level than an entire institution. This demonstrator will use the discipline-cluster common data model described

above and apply a meta-schema across eCrystals, Materials Data Centre and the new nanofabrication centre repositories. This will provide an orthogonal view, compared to the conventional discipline-specific ones.

Conclusions

The Institutional Data Management Blueprint project has carried out a data management audit across the School of Chemistry, Electronics and Computer Science, Engineering Sciences and Humanities using both top-down and bottom-up approaches. We have applied AIDA, questionnaires and numerous interviews to obtain both qualitative and quantitative picture of the current state of research data management. It is apparent that there is room for improvement to develop a coherent data management approach, as the current business model is not scalable, nor sustainable, to meet the current and future demand required to support the university's strategic goals[7].

A three-layer metadata strategy based on Dublin Core has been proposed to provide a unified approach to improving data management across all disciplines. Three pilot implementations around archaeology, the nano-fabrication centre, and meta-search across federated repositories, have been described and development work is starting on these.

Specific quick-wins that we have identified include: a one-stop shop for research data management, piloting an institutional data repository and develop a scalable business model. In the longer term we aim to embed research data management training, provide comprehensive backup in a seamless way for all, help researchers manage their data throughout its lifecycle, and look at an open data mandate.

IDMB has succeeded in taking a holistic approach to research data management at the institutional level, with a team comprising academics, IT specialists, librarians, research support services and senior management. By working together we are confident that rapid progress and long-term strategy, based around a sustainable business model, will be achieved.

It is clear that the current data management situation at the University of Southampton is analogous to the HPC landscape at Southampton a decade ago. The institution successfully moved to a more coordinated HPC framework since then that provides world-leading capability to researchers through a sustainable business model. A similar step change in data management capability is required in order to support researchers to achieve the University's ambitious strategic aims in the coming decade.

Acknowledgements:

The authors thank JISC for funding this work. We also thank Graham Pryor (DCC) and Sally Rumsey (Oxford University) for their contributions as part of the project steering group.

[7] http://www.soton.ac.uk/strategy/