The International Journal of Digital Curation
Issue 2, Volume 3 | 2008

The Data Audit Framework: A First Step in the Data Management Challenge

Sarah Jones,
DAF Project Manager,
Digital Curation Centre, HATII at University of Glasgow

Alexander Ball,
Research Officer,
UKOLN, University of Bath

Çuna Ekmekcioglu,
Information Services, University of Edinburgh

November 2008

Summary

The Data Audit Framework provides organisations with the means to identify, locate and assess the current management of their research data assets. Armed with this information they are in a position to improve ongoing data management. In this article we share our experiences of implementing the Framework and report back on the kind of data issues researchers commonly face. We also indicate how the Framework will be further developed before being released for widespread adoption.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Overview of the Data Audit Framework

Although vast quantities of data are being created within higher education, few institutions have formal strategies in place for curating these research outputs in the long-term. Moreover there appears to be a lack of awareness as to what data are held and how they are being managed (Lyon, 2007). The JISC-funded Data Audit Framework (DAF) has been developed in response to these issues. If institutions are to be in a position to manage and share their data, they must first establish an overview of holdings and the policies and practices in place to manage them. The Data Audit Framework provides a mechanism for collecting such information through its audit methodology and supporting online toolkit.

Five projects were funded by the JISC through its repositories programme to complete this work: the Data Audit Framework Development (DAFD) project led by HATII at the University of Glasgow and four pilot implementations, which are being run at the University of Edinburgh, King's College London, Imperial College London and University College London. They will test the Framework in a range of contexts and report back on its applicability to the UK Higher Education research communities. Their work will also provide an opportunity to explore user expectations, in particular which benefits are most important for Higher Education communities so we can ensure the audit data delivers on these requirements.

Auditing data can bring several benefits for an organisation. They could be categorised into efficiency savings, risk management, and enabling access and reuse. Realising all of these benefits relies on knowledge of data holdings. Being aware of what is held and by whom can identify duplication of effort and enable prioritisation of resources. Knowing how data are being curated, and whether controls are in place, will point to areas of potential risk. Similarly, an understanding of data agreements is crucial to facilitate access and promote reuse. Thus, knowledge of holdings is the cornerstone of effective data management. The Data Audit Framework is a first step in this process, assisting organisations to collect such information so they can develop policies and processes appropriate to their needs.

The DAF Self-audit Methodology

The DAF methodology was conceived by Sarah Jones, Raivo Ruusalepp and Seamus Ross from HATII at the University of Glasgow. It was designed to be applied without dedicated or specialist staff. Subject-specific expertise is helpful but is not viewed as essential. An understanding of data issues and curation practices takes precedence. Staff with experience of managing data or with a qualification in library, archive or information management would be particularly suited to the role of auditor. Personal characteristics such as those suggested in ISO 19011 (2002) – open-mindedness, diplomacy, perceptiveness and self-reliance – are also key.

Figure 1. Stages in the DAF methodology, © 2008 HATII, University of Glasgow.

The audit methodology consists of four stages as seen in Figure 1. In the planning stage the purpose and scope of the audit are defined. Preliminary research is conducted and meetings scheduled so time spent with the organisation's staff can be optimised. The purpose of the second stage is to establish what data assets exist and to classify them according to their value to the organisation. The classification step determines the scope of further audit activities, as only the vital or most significant assets are assessed in greater detail in the following stage. The information collected in Stage 3 helps to identify weaknesses in data policy and current data creation and curation procedures. This provides the basis of recommendations in the final stage of the audit. The knowledge gained from the audit will enable the organisation to improve its data management policies and processes.

Validating the Methodology

The DAF methodology was tested in pilot audits run at development project partner institutions. Çuna Ekmekcioglu audited data assets in the School of GeoSciences at the University of Edinburgh, a leading international research centre; Alex Ball worked with the Innovative Design and Manufacturing Research Centre (IdMRC), a research group within the Department of Mechanical Engineering at the University of Bath; and Sarah Jones focused on Glasgow University Archaeological Research Division (GUARD), a commercial research unit within the Department of Archaeology. An explanation of how the audit methodology was implemented in each of these cases is presented below. The article will then discuss the common data issues encountered across each test case.

School of GeoSciences, University of Edinburgh

The University of Edinburgh is a research-led university. Research data are generated by individuals and research groups in all 21 Schools of the three Colleges. These span a very wide range of disciplines across the Humanities and Social Sciences, Medicine and Veterinary Medicine, and in Science and Engineering. The School of GeoSciences is one of the largest schools in the College of Science and Engineering and a leading international centre for research into GeoSciences, with some 80 academics, 70 research fellows and 130 PhD students. The School staff contribute to one or more of five Research Groups (Earth Subsurface Science, Global

Change, Human Geography, Edinburgh Earth Observatory, Centre for Environmental Change & Sustainability) and may be involved in inter-University Research Consortia and Research Centres.

Data Audit Framework methodology was tested in a pilot audit in the School through a series of semi-structured interviews with over 30 staff from the Global Change and Human Geography research groups. Implementing the methodology was straightforward. The main challenges were the time required to set up interviews, lack of documentation on data management practices and restricted access to shared network drives. Although the pilot was not a comprehensive audit, information collected provided a detailed view of the volume of data assets, data types, storage and back-up issues, current skills gaps in data management, and issues with the retention of the data assets. As such it was sufficient to provide useful recommendations on steps to improve data creation and management practices.

In the light of the experiences gained from the pilot audit, a further five audits are being conducted in the Institute of Astronomy, School of Molecular and Clinical Medicine, School of Biological Sciences, School of History, Classics and Archaeology and the School of Divinity. These audits are expected to be completed in early 2009.

IdMRC, University of Bath

The Innovative Design and Manufacturing Research Centre (IdMRC) is a research group within the Department of Mechanical Engineering at the University of Bath. It was set up in October 2001 with funding from the EPSRC's IMRC programme, and is one of 16 such centres in the UK. The Centre has 14 academics, three research fellows, 16 research officers and 20 research students. The IdMRC's work is widely supported by industry, especially from the aerospace and packaging sectors and with emerging strengths in shoe and electronics manufacture.

As the Centre already has an interest in knowledge and information management, the Director saw the benefits of performing the audit and agreed to take part. Due to strict security policies, access to the shared network drives was not granted, so only a limited amount of preparatory work was possible. As a result, information had to gathered by asking researchers in turn. This was accomplished first through a series of interviews, starting with the theme administrator for each of the Centre's four research teams. The interviews were wide-ranging, covering not only the identification and classification of data assets but also data management practices observed. The breadth of the interviews meant that later clarifications could be sought and provided by email, thus avoiding the need to schedule follow-up interviews. Snowball sampling was used to choose further interviewees, something that was possible due to researchers being on site and available for interview much of the time. Those researchers not interviewed (approximately two thirds) were invited to contribute by means of a questionnaire.

The audit revealed that data management issues were recognised and addressed within the Centre, although practices varied between themes and researchers. The final report of the audit proved to be a useful medium for sharing good practice across the Centre, as well as allowing the auditor to introduce further suggestions.

GUARD, University of Glasgow

The Glasgow University Archaeological Research Division is a commercial arm of the Department of Archeology at Glasgow. The Unit was founded in 1989 and currently has 33 members of staff. It offers a wide range of services from consultation to fieldwork and post-excavation analysis. Staff are constantly engaged in projects that produce digital data assets ranging from photographs and image collections, to computer-aided designs, GPS/GIS, and analysis data.

Implementing the methodology posed no major issues. The Director of GUARD was already aware of data issues within the Unit and so was keen to take part. Access was granted to the shared drives on which most data was held, therefore much of the preparatory work and identification could be done remotely. The main challenge during the audit was arranging times to meet with staff - much of the Unit's work is conducted off-site so staff availability was low. This was exacerbated by the audit taking place in the summer when many other staff were away on annual leave. Delays in setting up interviews increased the elapsed time needed. Interviews were arranged with around a quarter of the workforce. Some comprised general discussions on data curation practices but most focused on specific data assets and were crucial in completing the assessment stage. The interviews proved very useful for understanding how staff created and managed data and enabled the auditor to identify areas for improvement. Staff were forthcoming with suggestions of changes they felt might enhance digital curation practices. These aspects helped feed into recommendations as to how data management could be enhanced.

Common Data Issues

The issues researchers face in terms of creating and maintaining their digital research data appear to be shared across the disciplines. Similar data issues were encountered in all three pilot audits despite the differing institutional and research contexts. The main issues encountered centred on storage provision, lack of data policy and issues with legacy data. They are explained in more detail below.

Storage Issues

When discussing data needs with researchers, storage was often noted as inadequate. In several cases we found researchers resorting to storing their data locally, on personal PCs at home, or on external storage devices such as hard drives, data sticks or CDs. When asked about maintenance, practices were generally ad hoc; very few researchers adopted a robust approach to backup despite being well aware of the risks faced. Examples were provided of data loss and irretrievability due to poor maintenance. In one of the pilot audits a researcher sadly pointed out that results of 40 years of research could no longer be accessed as the data were initially stored on CDs which had become corrupted.

In some cases additional storage was available, so many of these issues could be avoided. At the University of Edinburgh additional capacity is generally provided on request. Few researchers however appeared aware of this service. A large contributing factor to GUARD's data storage is the need to maintain legacy data as projects may re-open. As such, investigating alternative storage for inactive data may help alleviate storage pressures in the meantime. At the University of Bath, the use of the shared network drives was patchy, partly due to a lack of clarity over what each drive should be used to store.

Where the use of non-networked, centrally supported storage is unavoidable, basic procedures such as integrity checking and optical media refreshment should be established to minimise the risk of data loss. Maintaining an accurate locations register, and ensuring naming conventions and filing structures are transparent to enable future users to locate and interpret the data, will also be crucial.

Data Policies

Across the pilot audits we encountered a lack of formal policies for creating and managing data. GUARD has well-defined practices for paper records and archaeological finds, but while some of them had transferred over to their digital collections, the prevailing approach was very idiosyncratic and largely defined by the individual researcher. This held true in the case of GeoSciences and the IdMRC too. Pockets of good practice were mixed with less sustainable approaches. In general there was a lack of standardisation in terms of version control, file naming conventions and directory structures. These factors made it difficult for researchers to work collaboratively as it was not always clear where to find data and which was the most accurate version to use. The majority of researchers had first-hand experience of these issues and so were keen to obtain guidance on improving data creation to limit such effects.

Various suggestions were provided in the final audit reports to help organisations control the data creation process to enable more effective data management and reuse. In terms of file naming, various conventions could be used ranging from replacing spaces with underscores and differentiating words in a string by capitalising the first letter, through to adopting the 8.3 file naming standard to assure interoperability across platforms. Guidance on various options can be found through advisory services such as TASI (Technical Advisory Service for Images, 2006). Data policies do not need to be complex - indeed, those excessively detailed and rigid may be too difficult to implement - rather they should provide basic guidance on procedures to be followed to ensure consistency in approach. Data issues are not unique to departments, so an institution-wide approach, perhaps developing a broad top-level policy that could form the basis for departmental work, would be a fruitful beginning.

Legacy Data

A place of deposit is not always provided for the long-term preservation of researchers' data. While several research councils provide this service, such as the ESRC through the UK Data Archive (UKDA), and NERC through its subject-specific data centres, data management and preservation are often the concern of the individual researchers and their institutions. As they are not always equipped to fulfil this role it poses issues for the data in the long term. In the pilot audits we regularly encountered a lack of data controls, such as access restrictions and edit rights, which had the potential to lead to data corruption and integrity issues. Researchers were at a loss as to how they should maintain data in the long term; a common approach to legacy collections was benign neglect. The problems caused by a lack of active management were exacerbated when researchers moved on to new roles. With inadequately documented data, there was a real danger of the significance of the data being lost to the organisation. Furthermore, not having a person assigned as responsible for the data meant organisations were unaware of what existed or seemed unclear on how they could gain access to something they knew they held.

As deposit is not always an option, organisations need to engage with long-term data management to support their researchers who are struggling to achieve this. A preliminary step would be to register data, noting its location and any restrictions such as confidentiality conditions to ensure holdings can be monitored and that data do not become "orphaned" within the organisation. Depending on the organisation's resources and data issues, basic processes such as integrity checking and optical media refreshment could be implemented to avoid data loss. Moving legacy data to alternative storage with limited access could also minimise the risk of inadvertent data corruption; something which might otherwise go unnoticed since the files would not be in active use.

Recommendations

The two main recommendations to come out of the initial pilot audits are that data policies and training for researchers are urgently required.

1. An institutional data policy with guidance on best practice in data creation, management and long-term preservation would provide departments with a basis from which they could develop policies and practices suited to their particular contexts. Researchers are calling for central guidance as they have experienced the problems poor data creation and management can bring. If approaches are standardised across organisations, many of the issues faced will be minimised. Creating policies cannot be a standalone action; in order to implement them, researchers need support and training.

2. There are many sources of advice and best practice available, for example from services such as the Digital Curation Centre (DCC), the Digital Preservation Coalition (DPC) and data centres like UKDA and the former Arts and Humanities Data Service (AHDS). These services, along with local providers such as the archives and library, are also likely to provide training on creating and managing data. Ideally such training would take place early in the research process or career of the researcher. Equipping postgraduates with basic information management skills and awareness of data issues should help embed a more robust approach to data creation and management. In addition, providing expert support to enable researchers to produce sound data management plans while applying for research grants, and encouraging applicants to include a budget to cover data management, will help raise awareness of the issues. It will also ensure involvement of information professionals from the start of the research process.

The Data Audit Framework provides a starting point, helping organisations take stock of the current state of play. Auditing identifies organisations with an overview of the current state of play. It identifies the main data issues researchers face, recognises areas where data are at risk, and helps to plan for future infrastructure requirements. Once aware of their data issues and needs, organisations can then plan work to safeguard their data assets for the future.

Further Work

Work on the Data Audit Framework continues through the four JISC-funded implementation projects previously mentioned. The four projects are conducting some 20 audits across a range of subject areas in research groups, departments and schools of various sizes with a diverse range of data collections. They are expected to report in

late 2008 or early 2009. In addition, the Universities of Oxford and Southampton have also started conducting DAF audits to tie in with their respective Research Data Management and DataShare projects. The findings from all these initial users will provide valuable feedback on the applicability of the methodology in various contexts and the usability of the online tool. Recommendations will be followed up by the development team to enhance the Framework.

Conclusions

The Data Audit Framework promises to be a valuable tool for organisations seeking to make the most out of their digital assets, providing benefits such as efficiency savings, risk management, and improved access and reuse. The three initial pilot studies revealed that researchers require basic training and guidance on creating and managing their digital assets. The issues faced, such as lack of storage, irretrievability and data loss, were common across discipline and institutional boundaries. We encountered several calls for high-level guidance, suggesting that the creation of institutional data policies would be a useful first step in tackling data management. The issues researchers faced were exacerbated by the fragmented infrastructure for long-term data preservation. Only a limited proportion of the data being created was served by a specialist data centre or repository. In many cases long-term preservation was the concern of individual departments or institutions, which at present rarely have the capacity and skills to handle their data. If we are to safeguard our data until a fully equipped network of repositories is in place across the UK, investing in data management training and support for researchers is essential.

Outputs from the DAFD project are free to use and available online[1]. The DAF methodology and online tool are currently being tested by four implementation projects and will be updated in light of their feedback. The final versions will be available through the project website in early 2009.

Acknowledgements

The authors of this report would like to thank the the Joint Information Services Committee (JISC) who made this research possible through a grant from their Repositories Programme. Thanks also go to colleagues at HATII, Edinburgh, UKOLN and King's College for their assistance and feedback during the project and to the DAFD steering committee for its guidance throughout. We are hugely indebted to the School of GeoSciences, Edinburgh, IdMRC in Bath and GUARD in Glasgow for agreeing to take part in the project and allowing us to use them as test cases for our research.

References

ISO 19011:2002. Guidelines for quality and/or environmental management systems auditing.

Lyon, L. (2007). *Dealing with Data: Roles, Rights, Responsibilities and Relationships.* Retrieved October 30, 2008, from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

[1] Data Audit Framework (DAF) http://www.data-audit.eu

Technical Advisory Service for Images. (2006). *File naming*. (Advice Paper). Retrieved October 30, 2008, from http://www.tasi.ac.uk/advice/creating/pdf/filenaming.pdf