Project Information			
Project Acronym	DAFD		
Project Title	Data Audit Framework Development (DAFD) Project		
Start Date	01/04/2008	**End Date**	30/09/2008 extended to Dec 2008
Lead Institution	University of Glasgow		
Project Director	Professor Seamus Ross		
Project Manager & contact details	Sarah Jones, s.jones@hatii.arts.gla.ac.uk +44 (0)141 330 3549		
Partner Institutions	University of Edinburgh; King's College London, UKOLN, University of Bath;		
Project Web URL	http://www.data-audit.eu		
Programme Name (and number)	Digital repositories programme 2007-8		
Programme Manager	Neil Jacobs		

Document Name			
Document Title	DAFD Final Report		
Reporting Period	n/a		
Author(s) & project role	Sarah Jones, Project Manager		
Date	15/01/2009	**Filename**	DAFDfinalreport.doc
URL	n/a		
Access	√ Project and JISC internal		□ General dissemination

Document History		
Version	**Date**	**Comments**
1.0	23/10/2008	Original submission
1.1	15/01/09	Revised submission after online tool testing

Data Audit Framework Development (DAFD) Project

Final Report

Sarah Jones, Seamus Ross

15th January 2009

Contact: Sarah Jones s.jones@hatii.arts.gla.ac.uk

JISC Final Report

Table of Contents

Acknowledgements

The Data Audit Framework Development project was funded by the JISC through a grant from its Repositories Programme. Work was led by HATII at the University of Glasgow in collaboration with partners from the Estonian Business Archives, University of Edinburgh, UKOLN (University of Bath) and King's College London. We are grateful to these partners and the Steering Committee for their insights and detailed feedback on the applicability of the Framework. Thanks also go to the Digital Curation Centre which acted as an umbrella for this work. We acknowledge the support of the four pilot implementation projects at the University of Edinburgh, Imperial College, King's College London and University College London who will test the methodology and online tool to promote its uptake.

Executive Summary

Effective management and reuse of research data have become critical success factors for excellence in the education community. Many institutions however are unsure what data they hold or where to begin in terms of curation. The Data Audit Framework (DAF) has been developed in response to help bring order. It provides an audit methodology and supporting online tool that enable organisations to establish an overview of data holdings and the policies and practices in place to manage them. Adapted to the diverse needs of UK Higher Education communities, the tool is non-prescriptive and flexible. It has been tested in a series of pilot audits by the development team and continues to be trialled by four pilot implementation projects. The lessons learned throughout this process are being captured as practical guidance in the methodology and shared on training courses.

The approach to developing an audit methodology drew on lessons learned by the DRAMBORA team when creating a repository self-audit toolkit. The DAFD approach similarly began with research into user needs to determine what information should be captured. The draft methodology was then tested in a series of pilot audits and refined in light of feedback. The validated audit methodology suggests four incremental steps: 1) planning the audit, where scope and expectations are defined; 2) identifying and classifying assets, where an inventory is created and the data the organisation wants to consider in more detail is identified; 3) assessing the management of data assets to consider aspects such as rights, reuse, storage, backup and integrity checking; and finally 4) reporting and recommendations where risks are noted and steps to improve ongoing data creation and management are provided.

Similar data issues were encountered across the organisations in which DAF was piloted despite differing institutional and research contexts. The main issues researchers faced centred on storage provision, lack of policy and legacy data. Inadequate storage meant data was often kept on external devices causing problems such as irretrievability, poor backup and carrier vulnerability. A lack of policy meant idiosyncratic working practices prevailed, leading to differences in naming conventions, filing structures and versions control procedures, hindering data sharing and collaboration. Finally, a general lack of procedures to control access and edit rights meant the integrity of legacy data was at risk. While the findings of the initial pilot audits are not sufficient to draw full conclusions, discussions with the four implementation projects suggest our results are indicative of wider trends.

The findings from our research have helped inform other data projects such as the UKRDS and data skills study. The Framework itself has been the subject of numerous well-received presentations, research papers and training sessions. Although it is still early to evaluate impact, responses seem very positive. There are also several opportunities for continued development of the online tool and a growing community of potential users established through outreach activities.

Background

One of the current challenges for Higher Education Institutions in the UK is their efficient participation in the national knowledge economy. Management and reuse of research data have become critical success factors for excellence. The recent Report of the OSI e-Infrastructure Working Group presses this agenda, pushing for a national e-Infrastructure to enhance the global standing of UK research.[1] While research data offer benefits, they also pose risks; reaping the benefits while managing these risks requires knowledge of data holdings. If institutions are to manage and exploit their research data, they must be able to quickly and easily establish an overview of collections and the policies and practices in place to manage them. This need was identified by Liz Lyon in the seminal JISC-commissioned *Dealing with Data* report, which recommended:

> a framework must be conceived to enable all Universities and colleges to carry out an audit of departmental data collections, awareness, policies and practice for data curation and preservation.[2]

The Data Audit Framework Development (DAFD) project was conceived in direct response. As such it sought to provide a mechanism and online tool to collate, manage and share information on research collections and data management practices. Knowledge of data holdings is a prerequisite for their effective management. As data audits provide this information they can bring several benefits for an organisation. These benefits could be categorised into efficiency savings, improved risk management, and enabling access and reuse. The initial test audits have shown that different benefits act as the impetus for an audit in each context. At the University of Edinburgh, for example, IT managers were interested in finding out what was held on the server to identify duplicates, while GUARD became involved to improve archiving workflows. From the promotion and outreach work we've done it seems the data librarian and curator communities will particularly benefit from this tool, both in capacity planning exercises and to improve ingest workflows and submissions to Institutional Repositories.

The project was led by HATII at the University of Glasgow in collaboration with the DCC. As such the team were in a position to draw on lessons learned when developing similar tools like DRAMBORA, and able to access a wealth of experience in the field of digital curation. The DAFD team set out to develop a general solution that meet the diverse needs of the HE/FE community, but also recognise differences across research field and organisation type. This was achieved by conducting a series of pilot audits in various research areas. In addition the development team worked alongside the four DAF pilot implementation projects funded by the JISC to take forward the outputs of the project and embed them in research environments.

Aims and Objectives

The aim of the project was to develop a Data Audit Framework (DAF) adapted to the current needs of UK Higher Education institutions. The project set out to create an audit methodology and a software tool intended to support and facilitate data audit. The toolkit should include a registry component to facilitate the recording and cross-institutional data sharing of audit results.

Specific objectives were to:

- develop a self-audit methodology suited to a diverse range of data and organisational types;
- validate the methodology through pilot audits and make necessary revisions;
- develop an online tool and registry guided by lessons learned on pilot audits;
- test and redevelop the online tool to ensure it is fit for purpose;
- establish links with pilot implementation projects and support their use of the Framework;
- promote the Framework to raise awareness of its potential and encourage uptake.

[1] OSI e-Infrastructure Working Group, *Developing the UK's e-infrastructure for science and innovation*, London, 2007, available at: http://www.nesc.ac.uk/documents/OSI/report.pdf

[2] Lyon, Liz, *Dealing with Data: Roles, Rights, Responsibilities and Relationships,* Bath, 2007, available at: http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

Methodology

Project approach

The timeframe for the project was quite short and outputs were needed quickly to allow the pilot implementation projects to begin work. We set out to create a validated methodology in the initial two months so effort could focus on developing an online audit tool and guidance to support those using the Framework. The project began with research. Potential audit approaches and the needs of UK Higher Education were investigated to define a suitable approach to data audit. This was validated through pilot audits and refined in light of feedback. Each development partner was required to conduct an audit within their institution. These were to take place in diverse subject areas and different sized departments or research groups. This approach was chosen as it was found to be successful during the DRAMBORA project, which created a repository self-audit methodology and validated it through several audits based in very different organisations across Europe.

The completed audit forms and validated methodology provided a stable input from which to develop the online tool. A list of requirements was also compiled from auditor feedback to guide development. Creating documentation such as this was crucial throughout the project to ensure the development process could be tracked and examined. The tool development process drew on previous experience HATII developers have in this field. The technologies used (PHP and MySQL) were selected on the basis of developer familiarity and their open source credentials.

Audit methodology

In terms of the audit methodology, guiding development factors were simplicity and applicability. We intended to keep the process short and streamlined so the time investment needed to understand and implement the methodology was minimal. As the methodology needed to fit diverse research contexts and data types the approach was non-prescriptive, instead offering various alternatives for completing steps. Information on data assets, for example, could be collected by questionnaires, interviews, desk-based research or observational studies. Moreover the element set was designed to be generic. Using Dublin Core as a base, we altered some field names and descriptions to better suit data and added some new fields to capture information on data management. The broad applicability of Dublin Core has meant the information asked for is not too specific. Indeed, some users have requested the possibility to extend the element set to add more subject related information.

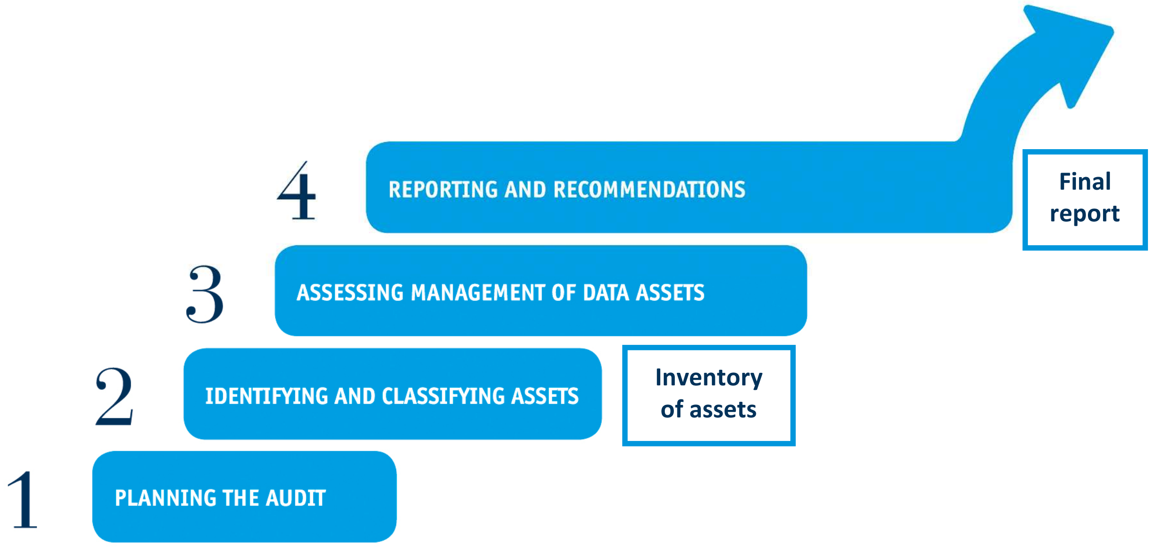
4
REPORTING AND RECOMMENDATIONS
Final report
3
ASSESSING MANAGEMENT OF DATA ASSETS
2
IDENTIFYING AND CLASSIFYING ASSETS
Inventory of assets
1
PLANNING THE AUDIT

Figure 1: The DAF Methodology and outputs

There are four stages to the audit methodology, as seen in Figure 1. Initially this had been a five stage process with identifying and classifying being two separate steps. Auditor feedback however suggested that it was easier to collect the information at once. In fact, over time we have found that it is easier to flow steps together, for example if an important data collection is mentioned in an early interview, information for the identification and assessment stages could be collected at the same time. It's preferable to think of the stages as overlapping processes rather than discrete phases. The audit generates two key outputs: an inventory of data assets, divided into groups according to their importance for the audit; and a report that incorporates recommendations on how data management could be improved.

In the planning stage the purpose and scope of the audit is defined. Preliminary research is conducted and meetings scheduled to optimise time spent with the organisation's staff. The purpose of the second state, identifying research data, is to establish what data assets exist and classify them to determine the scope of further audit activities. The classification will vary based on the premise of the audit. If an institution is conducting a capacity planning exercise to determine future storage and preservation needs, the datasets they decide to look at in more details may be determined on the basis of how they were funded and whether provision has already been made for long-term preservation. Alternatively, an organisation may classify its data in terms of usage as it believes regularly used collections require most resource investment. In Stage 3 a chosen subset of the data collections will be assessed in greater detail. The information collected will assist auditors to identify weaknesses in data policy and current data creation and curation procedures. This will provide the basis of recommendations in the final stage of the audit.

Implementation

Applying the project approach

The project approach was well defined in the proposal so we followed the work plan laid out there. Work began with research to develop an audit methodology, which was subsequently tested in pilot audits. These results fed into online tool development and testing. The key aspects throughout all stages of work were communications and documentation, as the project partners and stakeholders were spread across the UK. A project wiki was established at the outset and regular phone calls and physical meetings took place. Updates on progress were shared with the JISC and other stakeholders. The two main stages of development – creating the methodology and online tool – also followed this model. Regular calls and meetings were held to discuss development and gather feedback on progress. In each case, development was incremental and iterative. Information was also shared with the wider community at every opportunity - several presentations and conference papers were given as well as disseminating information through promotional materials (leaflet and briefing paper) and the project website. During the project we were mindful of capturing information that would be useful to a broader audience so it could be repurposed later: advice from the initial test audits was captured in the form of 'auditor tips' and lessons learned reports; common data issues were shared through an IJDC paper; and our expertise in using the Framework was passed on in training courses.

We only encountered a few challenges when implementing the project methodology, the main ones were related to timing. The nature of the project meant four other projects (the pilot implementations) were reliant on our outputs. As we were all funded in the same period and shared the same project timeframe there was urgency from the outset to create an audit methodology to allow their work to begin. Research and scoping was restricted to enable completion of a draft methodology within the first month of work. One month was also allocated to testing the methodology through pilot audits. It took longer than anticipated to obtain information and set up meetings with staff in each organisation acting as a pilot, so it proved impossible to complete the audits within this timeframe. Two were well underway and two were still in the planning stages at the review meeting on 30th May. Three were complete by the end of June / early July. To reduce the impact of the late completion of the initial audits on the pilot implementation projects, we collated lessons learned and shared these in a conference call on 20th June, meaning they had the methodology and some guidance on using it to be able to start work. We also attempted to absorb this overrun by shortening the time provided for specifying online tool system requirements and completing development. It wasn't possible to finalise and test the online tool before the official project end date of October, however regular updates and access to a working prototype has been provided since then.

Another aspect worthy of note is defining the benefits of audit. In most cases we found an advocate within the organisation to be audited and worked with them to encourage wider participation. The locus behind the audit varied by case: the Director of GUARD for example was keen to improve archiving workflows to prevent data remaining within the Unit unnecessarily, while IT support in GeoSciences wanted to understand what was being held on the server. An interesting point was raised by the Edinburgh pilot with regard to this, namely that it sometimes felt like the researchers were helping the DAF project staff by participating rather than vice-versa. This most likely results from the varied expectations with which an organisation chooses to undertake an audit, as certain aspects of the methodology will be more and less relevant to their needs. In the current methodology there is not much scope for auditors to customise the information collected at each stage, other than selecting between the core and extended element set in Stage 3. Future users may require more flexibility. Difficulties can be encountered when agreeing a participant for the audit. This occurred in one of the test cases. The timing was not appropriate for several departments approached so work couldn't begin in time for the methodology validation phase. In the development project this fortunately presented a new opportunity: the delayed start of the audit at King's College meant it could be used as an additional test of the online tool, which sat particularly well since they were co-ordinating the evaluation and testing work package.

Validating the audit methodology

When tested in the initial pilot audits, the methodology proved appropriate and easy to implement. The contexts in which is was tested were varied: Cuna Ekmekcioglu audited data assets in the School of GeoSciences at University of Edinburgh, a leading international centre for research; Alex Ball worked with the Innovative Design and Manufacturing Research Centre (IdMRC), a research group within the Department of Mechanical Engineering at University of Bath; Sarah Jones focused on Glasgow University Archaeological Research Division (GUARD), a commercial research unit within the Department of Archaeology at Glasgow University, and Stephen Grace worked with the Centre for Computing in the Humanities at King's College London. Despite differences in domain, organisation size and type of research being conducted, the lessons learned in each audit were aligned.

Some preliminary desk-based research was conducted in each, though more could be conducted offsite in the GUARD audit and CCH due to remote access being granted to data, which was not possible in the case of Edinburgh GeoSciences and IdMRC. Interviews were a crucial source of information in each audit and essential to gain an understanding of the organisation's approach to data creation and management. The main lessons learned were: 1) elapsed time can be significant, especially when you need to hold several interviews; 2) the timing of the approach and internal advocacy / management backing is critical to ensure staff give up time to contribute; 3) scoping the work and adopting the correct level of granularity can be a challenge – snowball surveying and sampling techniques were used given the extent of data produced in each case; and 4) that collecting information early where possible or merging audit stages improves workflow. A detailed explanation of each test case and feedback from the auditors is provided in Appendix B: lessons learned reports.

Outputs and Results

Core deliverables

The key outputs from the project are the methodology and online tool. The methodology proposes a four stage process for conducting a data audit. Various approaches for completing the work are suggested, so the most appropriate method for the context can be used. Such flexibility was central to development of the methodology as we intended it to be applicable across all Higher Education Institutions. The online tool facilitates audit by providing a central place to capture the information collected. It assists data collection through aspects such as the questionnaire wizard, which feeds responses straight back into the audit forms. Features such as a status bar and post-it style notes are also provided to help auditors monitor progress and keep on track.

Knowledgebase

As the methodology was tested in a series of pilot audits, detailed knowledge on data issues faced by researchers was collected. Common issued faced included: inadequate storage, causing researchers to keep data on external storage devices with limited backup; lack of policy, causing inconsistencies in naming conventions, version control issues and data duplication; and lack of expertise to curate data or no central place of deposit, meaning data became irretrievable, corrupt or was lost overtime. These issues were summarised for a paper in the International Journal of Digital Curation.[3] The findings on data creation and curation practices was also shared with stakeholders such as the UKRDS and DCC as it provided a feel for the kind of support services and training required. The experience gained from performing audits was also captured to share with the pilot implementation projects and has been passed on through training sessions.

Outreach and promotion

Dissemination was a large part of the project. Various presentations, papers and training sessions were provided. These are listed below. We also held a launch event on 1st October, immediately after the iPRES conference, to profile the project and raise awareness of its outputs. A mail shot targeting Vice Principals for research, library directors, heads of computing science and institutional repositories was sent in advance of the launch. Announcements were also made on various listservs and through representative bodies such as UUK, UCISA, SCONUL, RIN, SoA, RMS, DPC, and DPE.

Presentations and poster sessions

- *The Data Audit Framework Development (DAFD) Project* and practical exercise on benefits of audit, presented by Sarah Jones on 11th June 2008 at Delos Summer School, Tirrenia, Italy
  http://www.data-audit.eu/docs/DAFD_overview.pdf;
  http://www.data-audit.eu/docs/Data_Audit_Framework_exercise.pdf

- *DataShare & Data Audit Framework projects at Edinburgh,* presented by Robin Rice on 12th June 2008 at the Edinburgh eScience Collaborative Workshop, NeSC, Edinburgh & 13th June 2008 at the Research Data Management Workshop, SAID Business School, Oxford
  http://www.data-audit.eu/docs/RICE_DataShare_DAF.pdf

- *The Data Audit Framework: A toolkit to identify research assets and improve data management in research-led institutions,* presented by Sarah Jones at iPRES 2008 on 30th September 2008, British Library, London
  http://www.data-audit.eu/docs/jones_a37.pdf

[3] Jones, Sarah, Ekmekcioglu, Cuna, & Ball, Alexander, *The Data Audit Framework: a first step in the data management challenge,* in International Journal of Digital Curation, Vol 3, No.2, 2008, available at: http://www.ijdc.net/index.php/ijdc/article/viewFile/91/62

- *Developing an audit methodology suited to research data assets*, presented by Sarah Jones at the DAF launch on 1st October 2008, British Academy, London http://www.data-audit.eu/docs/DAF_launch_methodology_slides.pdf

- *Experiences gained from implementing the Data Audit Framework*, presented by Cuna Ekmekcioglu at the DAF launch on 1st October 2008, British Academy, London http://www.data-audit.eu/docs/DAF_launch_implementation_lessons.pdf

- *Tools for managing research data: DAF and DRAMBORA,* presented by Sarah Jones on 20th October 2008 at Institutional and National Services for Research Data Management, Oxford http://www.data-audit.eu/docs/DAF_DRAMBORA_oxford.pdf

- *The Data Audit Framework,* presented by Sarah Jones on 13th November 2008 at the e-Science Collaborative Workshop, Sheffield http://www.data-audit.eu/docs/DAF_eScience_Sheffield.pdf

- *Poster session and tool demonstration,* International Digital Curation Conference 2008, Edinburgh, 1-3 December 2008 http://www.data-audit.eu/docs/DAF-poster-web.gif

- *The Data Audit Framework: a data management toolkit,* presented by Sarah Jones on 8th December at CNI task force meeting, Washington DC, USA http://www.data-audit.eu/docs/DAF_CNI.pdf

- *Introduction to the Data Audit Framework,* presented by Sarah Jones on 16th December 2008 at the DCC Information Day, Glasgow http://www.data-audit.eu/docs/DAF_info_day.pdf

Papers

- Jones, Sarah, Ross, Seamus & Ruusalepp, Raivo, *The Data Audit Framework: a toolkit to identify research assets and improve data management in research led institutions,* in conference proceedings of 2008 iPRES conference, p225-231, available at: http://www.bl.uk/ipres2008/ipres2008-proceedings.pdf

- Ekmekcioglu, Cuna, Ball, Alex & Jones, Sarah, *The Data Audit Framework: a first step in the data management challenge,* in International Journal of Digital Curation, Vol 3, No.2, 2008, available at: http://www.ijdc.net/index.php/ijdc/article/viewFile/91/62

Training

- Data Audit Framework presentation and appraisal exercise, DC101 summer school, NeSC, Edinburgh, 7th October 2008 http://www.data-audit.eu/docs/DC101_DAF.pdf

- Data Audit Framework half day workshop on 1st December 2008 at the International Digital Curation Conference, materials available at: http://www.data-audit.eu/promotion.html

In addition please see:

Name	**Location**
DAF methodology	http://www.data-audit.eu/DAF_Methodology.pdf
online tool	http://www.data-audit.eu/tool/
promotional leaflet	http://www.data-audit.eu/docs/DAF_leaflet.pdf
briefing paper	http://www.data-audit.eu/docs/DAF_briefing_paper.pdf
project poster	http://www.data-audit.eu/docs/DAF-poster-web.gif

Access to outputs

The main project outputs are available online, either through the project wiki[4] if internal, or on the public website at the URLs noted above. Information disseminated about the project was targeted at several user communities, as can be seen from the range of materials provided. Internal communications and progress can be tracked through minutes, updates and discussions on the wiki. The public profile of the project can be seen through the core outputs, promotional materials and presentations. Outputs from ongoing outreach activities will be available through the project website at: http://www.data-audit.eu/promotion.html

Outcomes

The project set out to develop a Data Audit Framework (DAF) adapted to the current needs of UK Higher Education institutions. The methodology created has been tested in four diverse pilot audits and was found to suit each context. It continues to be tested through the pilot implementation projects, which will report towards the end of 2008 / early 2009. Six core objectives were agreed at the outset of the DAFD project. The project achievements have been mapped against these.

Develop a self-audit methodology suited to a diverse range of data and organisational types

When developing the audit methodology we were mindful of the different types of institutions and contexts in which it would be applied. As such we created a flexible approach that offered different options for completing steps rather than being prescriptive. Recommendations are made to help users implement the methodology and guidance has been added following the initial pilot audits. The initial testing phase confirmed the methodology was appropriate for a range of data and organisation types.

Some preliminary feedback from the pilot implementations suggests a greater degree of flexibility may be beneficial. It appears the motivation for conducting a data audit can vary greatly: in some cases institutions are more concerned with scoping the quantity and type of data being produced while others chose to focus on working practices. Allowing auditors greater flexibility in how the audit runs, for example by allowing them to select metadata elements to construct their own audits forms or expand parts of the tool to allow a more survey-based approach, may be desirable. Feedback from the pilot projects will help determine which avenues are most appropriate for continued development.

Validate the methodology through pilot audits run by partner institutions and make revisions

The methodology was tested in pilot audits held within the development partners' institutions. Cuna Ekmekcioglu audited data assets in the School of GeoSciences at Edinburgh; Alex Ball worked with the IdMRC, a research group within the Department of Mechanical Engineering at Bath; and Sarah Jones focused on GUARD, a commercial research unit within the Department of Archaeology at Glasgow. These audits validated the methodology and identified changes needed and areas requiring additional guidance. Changes have been made to the methodology throughout the project following feedback from the development team, pilot implementation projects and early adopters. These are reflected in the different versions available on the wiki.

Develop an online tool and registry, guided by lessons learned on pilot audits

The initial system requirements drew heavily on discussion between the creators of the methodology and the tool developers. The project aims and audit approach were explained to ensure these were replicated by the tool. Lessons learned during the initial test audits were shared and a list of requirements drawn up to help guide development. The developers also benefitted from attending a Steering Committee meeting and conference calls so they could gather feedback from all project partners. As concerns about data sharing were raised in the initial test audits, a decision was made to remove the registry component from tool development. It was unclear what purpose it would serve and whether it would be used since institutions seemed more concerned with securing information than sharing it. This change opens up the possibility for greater flexibility in how the tool is used.

[4] See: http://wiki.arts.gla.ac.uk/dafd/index.php/Main_Page Please note this is password protected

Test and redevelop the online tool to ensure it is fit for purpose
The online tool testing was completed in early 2009. The delay was due to a late start on development as initial audits took longer than anticipated, and subsequent staff availability issues at King's College. A stable prototype tool was made available from October to limit the effect late testing had on pilot implementation projects. The technical errors noted were minimal, such as removal of blank fields and double-quotation handling errors. These have all been resolved. King's College conducted their development project audit during the online tool test. As such, a number of non-technical feature requests have also been provided based on practical user experience of the tool. These have proved very useful. Suggestions include providing a basic guide and FAQ on the home page, integrating DAF documentation through web links in the tool, and adding a completed 'best practice' audit that can be viewed by all users to show how the forms can be completed. It should be possible to complete basic improvements, such as the user guide, as part of ongoing DCC sustainability tasks. More in depth enhancements and redevelopment tasks are outwith the initial project scope.

Establish formal links with pilot implementation projects and support their use of the DAF
Contact was made very early on with the pilot implementation projects. A wiki was established in April to capture key records and project discussions. The implementations were provided with access to the wiki shortly thereafter. The first main discussion was held on 20th June when an introduction was given to the methodology along with an overview of lessons learned and tips on implementing the Framework. Cuna Ekmekcioglu also had follow-up discussions with staff from other pilot projects (Neil Jerrome from Imperial College and Panayiota Polydoratou from UCL) given her role on both the development team and as project manager of the Edinburgh pilot. Regular updates on progress and the online tool were provided via email, the wiki and listserv postings. The attendance of staff from the pilots at the DAF launch allowed these links to be fostered and additional guidance to be provided.

Promote the Framework to raise awareness of its potential and encourage uptake
The outputs section demonstrates the extent to which the Framework has been promoted. Several presentations, poster sessions and research papers have been given, both throughout the UK and internationally. As such we have raised awareness of the project across UK Higher Education and beyond. The response at events has been overwhelmingly positive and subsequent discussions have been very fruitful. The indirect impact of these outreach activities is also starting to be seen through word of mouth. We've recently been contacted by a data librarian from the University of Oregon who's read the methodology and IJDC paper and is now keen to start using the tool.

Several useful suggestions have been provided at events. One attendee at the launch recommended that DAF could be embedded in research workflows by allowing completion of certain elements to act as triggers to introduce other university services. Potential collaborators and users have also been found. At both the launch event and the IDCC, we've discussed how the DAF online tool can be used by BADC in their ingest and curation workflows. The raw code has been provided to allow them to investigate how components can be reused. Contact has also been made with potential collaborators outside of the UK, such as the Australian National Data Service (ANDS) and University of Kansas. We're currently pursuing opportunities to collaborate further with both organisations, particularly in the area of online tool enhancements.

Training has also been provided to pass on the expertise we gained when conducting audits. The response from our first workshop at IDCC was very positive – all participants noted something they could take back from the event to implement in their own institutions. These ranged from general concepts and advice on conducting data audits, to ideas for a local adaptation of the Framework.

The primary outcome of the DAF development project is increased awareness of data issues faced by researchers. Continued use of the tool will help identify common needs and infrastructure gaps. The Data Audit Framework is not only a first step for organisations wanting to improve their data management, rather its findings provide the basis from which to development of a UK data strategy. Key stakeholders to benefit from this research will be university management, Institutional Repositories and national bodies that support and represent Higher Education research.

Conclusions

The premise with which we set out on the project, namely that organisations are unaware of the data they hold and what policies are in place for managing them, stands true. In the initial audits we found a lack of data creation and management policies: practices were very ad hoc and depended largely on the individual researcher. This lack of policy was in spite of concerns from researchers and calls for central guidance. We encountered several cases where researchers wanted to reuse their data or share it with colleagues but were at a loss as to how to achieve this given their lack of expertise in data curation and limited resources. Many researchers hoped for an institution-wide stance on roles and responsibilities with regard to curation. As such, we recommend institutions create preservation policies that can be used as a basis for departments to establish local policies and procedures.

Data issues were common across the different contexts in which DAF was piloted. Lack of storage mean data was often kept locally or on external devices such hard drives, CDs and data sticks. The problems that ensued such as irretrievability, poor backup and carrier vulnerability, were a recurrent issue. Idiosyncratic working practices within small organisations also led to differences in naming conventions, filing structures and versions control procedures, making it difficult to share data and work collaboratively. Very few procedures were in place to control access and edit rights so the integrity of data over time was also at risk. As a large proportion of data created within UK Higher Education resides within departments in which few staff have data curation skills, there is a clear need for basic support and advice. Such support is essential if we are to safeguard our data until an adequate network of UK data repositories is in place to ensure their long-term preservation.

While the findings of four pilot audits is not sufficient to draw full conclusions, discussions with the implementation projects suggest our results are indicative of wider trends. Vast quantities of data are being produced, yet both basic training for researchers and a skilled workforce with the necessary infrastructure to curate this data is lacking. Roles and responsibilities for data curation are also unclear, as noted in the recent Key Perspectives study.[5] There is a huge need for training at all stages of the curation lifecycle, particularly basic data creation and management training for researchers and more intensive practical courses for specialist curators.

There are also several opportunities for ongoing development of the Data Audit Framework and online tool. These respond to feedback from auditors at our partner organisations, the pilot implementation teams and potential users who have become awareness of the Framework through presentations, workshops and our project website. Suggestions for future work are made in the following section.

Implications

The DAFD project has provided an initial approach to auditing research data holdings and investigating data management practices within HE institutions. This approach will be tested further through the four pilot implementation projects. The lessons learned on these projects will hopefully be gathered together to provide additional guidance on implementation and to act on recommended changes to enhance the Framework. DAF provides an inroad to help organisations gather information on their data issues and the kind of support services and infrastructure their researchers need. As such there is significant scope for HE communities to use the Framework as a tool to address their data needs and inform wider data strategy work being driven by bodies such as the JISC.

Opportunities for future development

Several ongoing development opportunities have arisen during the project to enhance the Framework to ensure it remains fit for purpose. These are outlined below.

[5] Swan, Alma & Brown, Sheridan, *The skills role and career structure of data scientists and librarians,* (2008), available at: http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf

- **Building on the lessons learned by the pilot implementation projects**
  The findings of the four pilot implementation projects being run at University of Edinburgh, King's College London, Imperial College London and University College London should be synthesised and their lessons learned fed back into the Framework. They are due to complete in late 2008 / early 2009. Agreeing additional development effort in Spring 2009 would ensure a full return on the investment in the pilot projects was realised. Activities could include: capturing lessons on various approaches to implementation in the form of practical user guides and training sessions; addressing redevelopment suggestions for the online tool, for example making it extensible and improving flexibility; and reviewing the methodology and audit stages in light of their wider application. By addressing the issues raised by the pilot projects early on we can ensure wider uptake is successful.

- **Redevelopment of the online tool**
  As the DAFD project was only six months long the time available for developing the online tool, including specification and testing phases, was limited to three months. This restricted the amount of additional functionality we were able to offer. Work focused on providing auditors with the required forms and basic functions such as the questionnaire wizard, to be able to collect and record relevant information on research data collections. There are several additions that could be made to the online tool during any redevelopment phase. These include: allowing audit forms to be extended and/or customised; building automatic recommendations into the tool so recording particular data types or formats flags up potential risks or links to best practice guides; allowing the audit to be saved and finalised at chosen points to provide a snapshot function similar to DRAMBORA; and improving the analysis tools provided at the final stage, for example by writing MySQL query plug-ins. The pilot implementation projects may also request additional features we have not yet considered.

  It would also be beneficial to add much more extensive online tool support as suggested in the testing conducted by King's College. This could include a manual, FAQ sheet, integrating the methodology guidance in the tool through web links, or providing an online demonstration or best practice audit to show potential users how the tool works. Some of these activities could be completed under the auspices of DCC as part of sustainability, though larger redevelopment tasks would fall outside this remit.

- **Continued outreach, training and fostering collaboration**
  Some promotion activities continue through the DCC as part of the project's sustainability plan. Initial workshops have been run as part of the DC101 curation summer school and at the IDCC conference. A session on using the Data Audit Framework will be run at the next DC101 course in March 2009 and further workshops will be provided on request. A continued phase of DAF activities would allow more promotion and training opportunities to be pursued. In addition we could provide greater effort and resources to assist those intending to use the Framework by developing ongoing partnerships. Various discussions have taken place with potential collaborators about the opportunities to pilot DAF in other contexts that are currently being pursued. These are noted in the 'outcomes' section of this report.

- **Data quality assessment tool**
  In the initial stages of the DAFD project, scope was restricted to identifying data holdings and investigating their management. When deciding how to maximise the potential of digital materials however, an understanding of data quality is needed. Without a clear indication of the value, accuracy and usefulness of research data it will be difficult to effectively promote reuse to enhance an organisation's reputation. A clear opportunity exists to build on DAF to address these issues. Aspects covered by the Framework such as the context of creation, documentation practices and data integrity, underpin issues of quality. Our audit approach and information collected naturally lead into a quality assessment. Research has already been undertaken in the field of defining metrics for data quality. We could build on this to create a quality component that would significantly enhance the DAF tool.

References

- OSI e-Infrastructure Working Group, Developing the UK's e-infrastructure for science and innovation, London, 2007, available at: http://www.nesc.ac.uk/documents/OSI/report.pdf

- Lyon, Liz, Dealing with Data: Roles, Rights, Responsibilities and Relationships, Bath, 2007, available at: http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

- Swan, Alma & Brown, Sheridan, *The skills role and career structure of data scientists and librarians,* (2008), available at: http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf

Appendixes

Appendix A: Glossary

ANDS	Australian National Data Service *a project to develop a national strategy for data creation and management in Australia*
BADC	British Atmospheric Data Centre *the Natural Environment Research Council's data centre for atmospheric sciences*
DAF	Data Audit Framework *a methodology and toolkit to audit research assets and improve data management*
DAFD	Data Audit Framework Development project *the project that developed DAF*
DCC	Digital Curation Centre *a UK-based centre to support expertise and practice in data curation and preservation*
DPC	Digital Preservation Coalition *A UK organisation to foster collaboration on the preservation of digital resources*
DPE	Digital Preservation Europe *a co-ordinating body for digital preservation research and best practice in Europe*
DRAMBORA	Digital Repository Audit Method Based on Risk Assessment *a repository audit tool developed by HATII with support from DPE and DCC*
GUARD	Glasgow University Archaeological Research Division *a commercial archaeology unit which agreed to be a pilot audit for the DAF*
HATII	Humanities Advanced Technology and Information institute *a research centre specialising in humanities computing and digital preservation*
HE / FE	Higher Education / Further Education
IDCC	International Digital Curation Conference
IdMRC	Innovative Design and Manufacturing Research Centre (IdMRC) *a research group in Department of Mechanical Engineering at University of Bath*
MySQL	*open source database management system using SQL (Structured Query Language)*
PHP	*open source scripting language to be used to create the DAF online tool*
RIN	Research Information Network *a UK organisation to promote the information needs of researchers*
RMS	Records Management Society *a professional association for all those who work in information management*
SCONUL	Society of College, National and University Libraries *a UK society to promote excellence in library services in HE and national libraries*
SoA	Society of Archivists *a professional body for UK archivists, archive conservators and records managers*
UCISA	Universities and Colleges Information Systems Association *a body to represent academic, management and administrative information systems*
UKRDS	UK Research Data Service *a feasibility study for a national shared digital research data service for UK HE*
UUK	Universities UK *the representative body for the executive heads of UK universities*

Appendix B: Lessons learned reports

These lessons learned reports provide feedback on the initial three implementations of the DAF methodology. The three audits were conducted within three different HE institutions in schools and departments of varying size and research area. Each took place between May-July 2008.

Engineering at University of Bath by Alex Ball

Background

In June and July 2008, an audit of research data assets held by the IdMRC at Bath University was conducted using the Data Audit Framework (DAF) methodology. This report briefly describes how the audit was carried out and the lessons learned from it.

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 sixteen such centres in the UK. It has four research themes: Advanced Machining Processes and Systems (AMPS), Constraint-Based Design and Optimization (CBDO), Design Information and Knowledge (DIAK), and Metrology and Assembly Systems and Technologies (MAST). 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.

Audit Stages

Stage 1: Planning the audit

An initial phone interview was held with the Director of the IdMRC to establish the scope, purpose and requirements for the audit. The IdMRC website was used to compile a list of staff and to clarify points about the history, structure and academic interests of the Centre; interviews were scheduled with the representatives (lead researchers) for the four research 'themes'. The IT Administrator was contacted about accessing shared drives.

Stages 2 & 3: Identifying and classifying data assets

A snowball sampling technique was used to choose interviewees, starting with the four theme representatives. In all, ten face-to-face interviews were conducted in this pattern. The interviews consisted of: 1. going through the interviewee's personal drives (and, where appropriate, shared drives) and determining which collections of data constituted data assets; 2. recording names, descriptions, statements of responsibility and locations; 3. discussing the importance of the asset in terms of current and future research; 4. recording additional information about file formats, software requirements, derived reports/ papers, dates of creation and update, etc.; 5. discussing how the interviewee managed the data. The remaining 17 members of the Centre were contacted by e-mail with a questionnaire similar to Audit Form2. This resulted in three completed questionnaires, two e-mail responses and one telephone interview. The resulting inventory consisted of 63 data sets, of which 18 are vital, 15 are important and 30 are minor. This was not comprehensive but was representative of the data assets of the Centre.

Stage 4: Assessing the management of data assets

Of the data assets in the inventory, 30 were chosen for analysis: the vital and important assets, less the three held by an external organization. Much of the information required for this stage had already been collected, with just a few gaps which were filled by e-mail queries. The basic metadata set was used in all cases.

Lessons learned

Time

Be prepared to 'badger' senior management. At various points the audit methodology calls for contact with the organization's management, who are among the busiest people in the organisation. A certain amount of persistence is required to keep the actual elapsed times in line with the estimates in the methodology. I found that e-mails to arrange phone calls worked better than long e-mails, and that phone calls were easier to arrange than face-to-face meetings.

Picking your moment. I found that attempting to perform the audit in June meant that academics were rather hard to get hold of (due to exams boards), but I had no trouble arranging meetings with research staff.

Choosing a sampling technique. I tried snowball sampling, starting with research 'theme' leaders. While I found it an effective way to get a good range of results, it did have a cost in terms of elapsed time, due to the fact that interviews were arranged on a rolling basis, instead of all at once, well in advance. I also tried filling in gaps with a questionnaire, but as predicted these had a low response rate.

Gathering information

Deciding on scope and granularity. The thing I had most trouble with was determining and communicating the scope of the audit. Very few of the research data sets held by the IdMRC were straightforward data bases or homogeneous data sets; in the main they tended to be ad hoc collections of resources supporting particular pieces of work. This made it hard to communicate the scope of the study in a way that would include, say, a set of a company's internal communications (used to analyse information flow within that company), while excluding collections of relevant literature accumulated by researchers in the course of their work. It also made it hard determining a suitable granularity at which to record the data assets. For example, two of the 'themes' are engaged in consultancy work, with each consultancy generating a small set of documents and data. On the one hand, each of these sets has its own character, life cycle and confidentiality conditions, and so could be treated individually, but on the other hand, taking them all together as one asset enables one to see better the overall character of the data and how the asset relates to other assets, and makes it considerably easier to make a comprehensive statement of holdings.

What information to ask for. I found that it saved more time than it wasted, to try and do as much information gathering in one go as possible. Instead of just getting a name, description and owner of a data asset in the initial interview, I also discussed the value of the data asset, and collected information such as creation dates, updating frequency, locations, formats and related publications. This meant I could just fill in any gaps later by means of a quick e-mail rather than having to reschedule an interview.

GeoSciences at University of Edinburgh by Cuna Ekmekcioglu

Background

In May 2008 an audit of data assets was conducted in the School of GeoSciences using the Data Audit Framework (DAF) methodology developed by DAF Development Team.

School of GeoSciences is a leading international centre for research into GeoSciences, with some 80 academics, 70 research fellows and 130 PhD students, and an annual research grant and contract income of around £4-6 million. In the last UK Research Assessment exercise, the School was rated as internationally competitive, receiving the top grade of 5/5* for its research. 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.

This report briefly describes how the audit was conducted and the lessons learned from conducting this audit.

Audit Stages

Stage 1: Planning the audit

This stage involved desk research such as browsing the School website, collecting annual reports and published articles, compiling a list of research active staff with their research responsibilities. Following on from this preparatory work, an initial meeting was arranged with the IT managers of the School to discuss how best the audit could be conducted and whether we could have access to the shared drives. Based on the information gathered from the desk research and the interview with the IT managers, the key staff were identified and invited to provide information about their research and research data via semi structured interviews.

Stages 2 & 3: Identifying and classifying data assets

In these 2 stages interviews were conducted with 35 academic/research staff, and an inventory of 25 data assets was created. The interviews were in the form of semi structured discussions to allow us gather as much information as we could such as data types, size of the collection, software used for analysis, value, storage, back-up, and retention of the data assets. Although this was not a comprehensive survey, the later interviews started to provide information already collected, suggesting the most important data assets had been recorded. Results of the pilot were reported back to the development team at a meeting in Glasgow at the end of May.

Stage 4: Assessing the management of data assets

Of the total 25 data assets only 4 of them were classified as vital assets and the detailed analysis of these assets was carried out buy the auditor contacting the owner of the data assets as and when necessary. It was rather difficult to complete all the metadata fields in audit form 3.

Stage 5: Reporting results and making recommendations

Generally speaking, the audit was useful to identify the gaps and issues in managing data assets in the School. Staff comments and suggestions for improvement of data management were found very useful. The results of the audit were drawn together and a final report was produced to recommend actions for change.

Lessons learned

Time

Time is one of the most important factors in conducting the audit successfully. If time was not a restriction we could have obtained better results for the audit. The planning stage should have been done well in advance, and the key staff should have been contacted at least a couple of weeks before the interview dates. In our case, most staff were out on field trips, or busy with marking exam papers and sitting in exam boards. Although 65 staff were contacted via email for interviews we could only interview 35 of them as the rest were not available for interviews until late June. The interviews themselves were time consuming. Ideally, an online survey could have been circulated to gather general information and then this could have been followed up by detailed interviews. As we did not have enough time to follow this approach we had to rely only on interviews.

Access to information

Access to the shared drives was denied on the basis of data protection. We could have accessed the drives if we contacted every research staff and obtain their written permission for access. However, this could have delayed the audit for another couple of months. Also, a great majority of the data assets was held on external hard disks, personal PCs and laptops, USB storage devices and CDs/DVDs. Since we could not access either the shared drives or data held on personal storage devices, the audit was conducted on project basis recording the data assets for the projects that were mentioned by staff in the interviews.

Scope and level of granularity

We had trouble with determining the scope and the level of granularity of the audit. We managed to identify only a couple of well described databases. The rest of the data assets we recorded were collection of text files, audio files, images etc. that were used in a particular research.

Documentation

We had difficulty with locating and accessing the documentation where data assets were described. There was no main index or catalogue of the data held in the shared drives – users were expected to search for the required data themselves, or with guidance from the relevant research staff. Searching for the data was difficult as most of the data was undocumented and there was not a well defined folder structure. Data was generally stored using a sub-folder file system where a variety of schemes (by year, by location, by relevant field trip etc.) were used. It was difficult to identify data from the file names only as there was no standard file conventions used while the data was stored.

Archaeology at University of Glasgow by Sarah Jones

Background

From May to July 2008 an audit of data holdings within GUARD at the University of Glasgow was conducted using the Data Audit Framework. This report considers how the methodology was implemented and reports back on lessons learned.

GUARD (Glasgow University Archaeological Research Division) is the archaeological practice within the Department of Archaeology. The unit was founded in 1989 and currently has 33 members of staff. It is a commercial provider offering a wide range of archaeological services, from consultation to fieldwork and post-excavation analysis. GUARD serves a range of clients including commercial developers, central and local government, public utilities and private individuals. Contracts are predominantly based in Scotland though staff also work in England, Ireland, Northern Ireland and further a field.

Audit Stages

Stage 1: Planning the audit

The initial planning stages were fairly smooth. The Director of GUARD was already aware of data issues within the Unit and keen to take part so it was not necessary to develop a full business plan. An initial meeting took place in early May to determine the scope of the audit and identify expectations. This gave the audit a clear focus and helped direct the initial stages of work. Access to the departmental shared drives was agreed which allowed research into the staff and work of the unit to be conducted. The information provided on the shared drive and website was ample to understand the context in which the Unit operated.

An email explaining the audit was sent to the Director to distribute to staff. This was thought sufficient, however when contacting staff during stages 2 and 3 they weren't always familiar with the work. In light of this a second interview was set up with the Unit's archivist to help gain some internal advocacy. This meeting proved very useful. The archivist was struggling with the Unit's data issues so was enthusiastic about the audit and introduced the auditor to staff members to encourage their participation.

Stages 2 & 3: Identifying and classifying data assets

A few interviews were requested to assist in identifying assets but the response rate was low. This was probably for a few reasons: the interview requests were generic rather then being targeted at specific areas of expertise or collections; staff availability was minimal as the Unit's work is largely conducted off-site; and people seemed unfamiliar with the audit. As access to the shared drives had been agreed however, the inventory could largely be created through desk-based research. The scope had been set as all work from the past three years excluding the forensic material due to sensitivities. After a preliminary survey a decision was made to record assets by project. GUARD conducts around 60 projects annually, each generally creating a small amount of data of a number of common types. As such, the inventory quickly became very large and entries were fairly consistent in terms of the type and size of data being created. In light of this a decision was made to complete a comprehensive survey for the first half of the scope and a sample for the 2005-06 period, picking out especially large, multi-stage projects or ones with unusual funding sources, data types or subjects.

This made the best use of time as effort was not spent replicating work yet still ensured the range of data assets was represented to allow all data issues to be investigated in the next stage.

The suggested classification was amended slightly to suit GUARD's data assets. While it could be argued that all projects are vital as the nature of archaeology means it would not be possible to reconstruct the data if lost, a classification was desirable to help decide which assets to analyse in greater detail in the next stage. It would have been unsuitable to categorise all active collections as vital as work was still ongoing on 42 collections from an inventory of 65. Instead classification was based on the revenue a project generated, since GUARD is a commercial unit, and whether it was still active or part of another piece of work. This acknowledged the trend for projects to be continued through subsequent pieces of work and the fact that the Unit was not responsible for long-term preservation as completed projects could be archived with RCAHMS.

Stage 4: Assessing the management of data assets
The detailed analysis of vital collections was conducted by interview. As much of audit form 3 as possible was completed in advance so this could simply be verified and enhanced in the interviews. Although it took some time to schedule the interviews, response rates were higher as most of the staff approached had already been introduced to the auditor and the requests were more specific, focusing on a particular data assets they had been involved in. Some general questions on how the member of staff created, managed and used data was asked at the start of the interview. This helped to build rapport and provided a useful overview of the Unit's approach to data that helped guide the interviews.

Stage 5: Reporting results and making recommendations
The interviews were very useful for seeing how the Unit created and managed data and identified areas for improvement. Staff were also very open with suggestions of what issues they faced and changes they felt necessary. These aspects helped feed into recommendations for change to improve workflows and minimise the risk of data loss and corruption.

Lessons learned
Timing is key – When investigating the organisational context it would be helpful to consider when is the best time to conduct the audit. Staff in GUARD are often out on fieldwork and as the audit was in the summer period annual leave also affected availability. It was anticipated the audit would be completed in May but delays in setting up interviews extended this to July, meaning it ran in parallel with other work the auditor was conducting, adding to the delays. Where possible an extended period of elapsed time where the auditor's diary is clear should be allowed. Otherwise planning should be moved forward to try to schedule main work in a short time period.

Scope the work carefully – The initial meeting with the Director of GUARD was very useful to scope the audit and identify expectations to ensure the audit delivered. The scope and level of granularity adopted should be flexible as it may be necessary to amend these during the audit.

Get internal advocacy – An email may not be sufficient to inform staff of the work and encourage participation. Attending a staff meeting, obtaining a personal introduction or securing some internal advocacy may be more successful methods. Involving core members of staff who are responsible for data management, such as the archivist, can help ensure the department accepts ownership of the audit results and take the recommendations on board.

Make best use of staff time - Agreeing access to internal documents is preferable as more of the research in the initial stages can be conducted off-site, thereby limit the demands placed on organisational staff in terms of questionnaires and interviews. It will also allow additional information to be collected in the early stages so time can be optimised in the interviews and discussion can focus on day-to-day working practices to see how data is being managed.

Centre for Computing in the Humanities at King's College, London by Stephen Grace

Background

Stephen Grace at the Centre for e-Research (CeRch) undertook a case study audit of the Centre for Computing in the Humanities (CCH) at King's College London (King's) during October-December 2008. He was helped by having the experience of the other audits undertaken in Glasgow, Edinburgh and Bath universities and thanks Sarah Jones, Cuna Ekmekcioglu and Alex Ball for their insights.

CeRch is a new research centre with a broad remit to work across discipline areas at the intersection between research methods and practice, digital informatics and e-infrastructure. It was established in October 2007 and launched in April 2008. It was based on the experience of the Arts and Humanities Data Service executive and the AHRC ICT Methods Network

Four approaches to other departments (in the School of Medicine) were made by Sheila Anderson, Director of CeRch and by Stephen Grace. These were unsuccessful for different reasons, and CCH was approached to help because of long-standing good relations with CeRch. CCH is a specialist research centre with an international reputation in the application of technology in research in the arts, humanities and social sciences. It has a teaching programme, but its primary focus is research activity which culminates in making digital resources available. The research projects, and their digital assets, are critical to the mission of CCH.

Audit stages

Each of the stages of the audit is taken in turn. The DAFD project decided to combine the Identifying and Classifying stages, and this case study treated these as a single stage.

Stage 1: Planning the audit

The original intention of the King's work package was to audit a department in the School of Medicine. One had been approached informally by Sheila Anderson before the DAFD project and expressed interest at that stage; it declined an invitation because of timing, as did another department. A third declined on the grounds that it "had no data" (it was a new research institute) and the fourth did not respond. This process was protracted and frustrating to the auditor and the wider DAFD project. CeRch was only recently established when the invitations were made, and its competence to assess data management may not have been clear to departments.

CCH was then approached in the person of the Research Fellow with overall responsibility for data management (DM), and he was willing to take part with the ready agreement of his Director. CCH is adjacent to CeRch and there are good relations between staff in the two centres. In addition, they share a System Administrator and this made the task of gaining access to servers very straightforward. User permissions were granted, and the auditor was able to gain access from his desktop within a few hours of agreeing the work with the Research Fellow. Travelling between departments was much reduced compared to the Edinburgh case study, for instance.

The Research Fellow was interviewed on data management practice and infrastructure (including plans and aspirations for the future). This gave the auditor a good general understanding of the distinctive culture of data management at CCH. A Research Associate was identified who could speak across a range of projects of his practical experience in managing research data.

Stage 2: Identifying and classifying data assets

Because of the delay in starting, it was decided to save time by scoping the data audit more narrowly than the whole department. At least fifty-six projects are listed on the CCH website, and the consistent DM practice identified at stage one suggested (as did feedback from the other DAFD case studies) that sampling would elicit enough evidence. Twelve projects were selected for audit, four each from the list of completed, stage two and current projects. A Research Associate with responsibility for data in four projects was interviewed about his data management practice.

CCH maintains an online list of its projects (overwhelmingly of curated digital assets), and this provided much evidence for Stage Two. Access to the CCH servers and directories made it simple to identify assets, but the process of collating and uploading information to the online tool is slow. CCH organises its servers into project directories and typically encompassing three sets of assets

- Digitised assets (images, digital texts, sound, etc)
- Marked-up copies of text in XML
- Files and scripts needed to render webpages

These hierarchies are established at the outset of a project, and helped the auditor in identifying and making sense of the data.

Assets were classified by project using the standard DAFD schema. Most were considered “vital” since the project websites were publically available or the resource was still being compiled.

Stage 3: Assessing the management of data assets

Because CCH has a coherent data management practice, it is easy to understand for each project where the digital assets exist and what forms they take. The Research Fellow establishes server and directory requirements at the outset, and user permissions limit the ability of an individual to alter standard practice in the centre. Researchers are aware of their responsibility to manage data, committed as they are to seeing the fruits of their work reach a wide audience. A second interview was held with the Research Fellow to gather more information and share preliminary assessments.

Much of the Assets register in Stage 3 (and Audit Form 3) was populated by the same sources of information used for Stage 2, especially the project descriptions on CCH’s website. These two stages may effectively be undertaken in parallel. It was time-consuming to enter the data online, and in future it may make sense to create a spreadsheet to hold the data compiled during an audit.

Stage 4: Final reporting and recommendations

The online tool generated a technical appendix and template for the final report which helped to reinforce the professional nature of the audit. These were delivered to CCH in January, and a debriefing interview was arranged.

Lessons learned

1. Identifying a department for audit

CeRch had four false starts before finding a willing audit partner. Partly this was because the centre itself was newly established, and maybe needed to establish its credibility across King’s. There was no central endorsement for the data audit project, which may have put the work in the context of other initiatives at King’s to improve infrastructure and support for research. A couple of departments found the initial timing of the audit in conflict with their work plans at the end of the academic year. The PI-DAF project at King’s will ensure that the benefits to the department are made explicit.

The final approach to CCH worked at least in part because the invitation was not sent in the first instance to the head of department but to someone known to have a role in data management. This approach will be used in the King’s PI-DAF project where known contacts in departments will be approached. With their support, the head of department will be asked to consent to the audit, approve any permissions (such as for server access) and confidentiality requirements. The DAFD publicity leaflet helped in explaining the audit process and the benefits of participating for researchers.

2. Organising time

There were no major problems with arranging interviews for the small cohort in CCH, although a wider schedule may have offered depth to the findings. The good relations between CCH and CeRch (and sharing a System Administrator) eased access to the private networks of CCH: granting this permission may be less willing, and involve more administrative burdens, in other departmental settings. It is critical to budget sufficient time, including lead time, in this as with arranging staff interviews.

Collating data and logging it on the online tool was time-consuming, even with a reduced sample of data. If the Data Audit Framework is to be widely used, it is essential every opportunity is made to speed up the process of the audit. This may be by importing data from a spreadsheet compiled by the auditor, or by delegating the collation task to others (see 4 below).

3. Need for documentation when using tool

The tool presumed the information is to hand when entering data for Stages 2 and 3. It would help if this data (collected from annual reports, documentation, web pages, etc) could be uploaded as a record of the evidence used by the auditor.

4. Availability of tool to manage audit process

The online tool is a useful way to manage the audit, and it could be enhanced to manage the full audit process. Interview dates, collation of survey information into Audit Forms 2 and 3, actions on recommendations, dates of reviews or follow-up audits could all be accommodated in a tool for an auditor (or team of auditors). In a devolved organisation like King's it is easy to imagine ways that the whole process may be undertaken by a range of actors – from a graduate student collating Form 2 in a single department, to a Records Manager overseeing the College's compliance with data security issues. Some of these issues are explored in the Scenario Test document compiled by CeRch for the DAFD team.