The Data Audit Framework: a toolkit to identify research assets and improve data management in research led institutions

Sarah Jones, Seamus Ross, Raivo Ruusalepp

Digital Curation Centre & Humanities Advanced Technology & Information Institute (HATII), 11 University Gardens, University of Glasgow, Glasgow, G12 8QJ
s.jones@hatii.arts.gla.ac.uk

Digital Curation Centre & Humanities Advanced Technology & Information Institute (HATII), 11 University Gardens, University of Glasgow, Glasgow, G12 8QJ
s.ross@hatii.arts.gla.ac.uk

Estonian Business Archives
Eesti Äriarhiiv Tartu,
Lembitu 6/8, 50406
Tartu, Estonia
raivo@eba.ee

At the time of writing the online toolkit was under development. It will have been tested ready for release before the iPres Conference. The Data Audit Framework will be officially launched on 1 October 2008 at the British Academy, London.

Abstract

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 exactly what data are held and whether they are being managed. In response to these concerns the Joint Information Systems Committee (JISC) issued a call for proposals to develop and implement a Data Audit Framework suited to the needs of the UK higher education research communities. The Data Audit Framework (DAF) Development project was funded to produce an audit methodology, online toolkit, and a registry. Four additional implementation projects were funded to test the toolkit and promote its uptake. This paper outlines the audit methodology, introduces the online toolkit, and provides feedback on implementing the Data Audit Framework.

Overview of Data Audit Framework

Project background

One of the current challenges for UK higher education (HE) institutions is their efficient participation in the national knowledge economy. Management and reuse of research data have become critical success factors for excellence in research. While research data offer benefits they also pose risks; reaping the benefits while managing these associated risks requires knowledge of data holdings. If HE institutions are to ensure they maximise their potential to exploit and reuse research data they must be able to quickly and easily establish an overview of the data collections they hold and the policies and practices that are in place to manage them. An audit framework offers a mechanism to collect, and manage such knowledge.

The need for an audit framework was identified by Liz Lyon in the JISC-commissioned report *Dealing with Data: Roles, Rights, Responsibilities and Relationships.* This report recommended a framework be conceived to:

> enable all universities and colleges to carry out an audit of departmental data collections, awareness, policies and practice for data curation[1]

The DAFD project team has produced such a framework. The methodology is simple yet flexible. As a result it can be applied across institutions irrespective of size, subject area or type of data created. A registry component will provide a mechanism to support the persistent recording of results of data audits based on DAF. This will allow organisations to share information on their data assets and curation policies while providing institutional and national perspectives to assist future data strategy development.

Project timescale

The Data Audit Framework Development project runs from April to September 2008 and is funded by the JISC under its JISC Repositories Programme.[2] Led by HATII at the University of Glasgow, the work is being conducted in collaboration with partners from the Estonian Business Archives, UKOLN at University of Bath, the University of Edinburgh, and King's College London. The project team has created an audit methodology and tested it in pilot audits that ran from May-July. Feedback from these audits enabled us to

[1] Lyon, L. 2007. *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, p5. The recent Report of the OSI e-Infrastructure Working Group presses a similar agenda if the UK is to ensure its research institutions adapt emerging e-infrastructure realities, see: OSI e-Infrastructure Working Group. 2007. *Developing the UK's e-infrastructure for science and innovation*, www.nesc.ac.uk/documents/OSI/report.pdf

[2] The total value of the Grant from the JISC is £ 100,000.

refine the methodology and has yielded information that is guiding the development of the online toolkit.

A beta-version of the online toolkit will be released in September 2008 to be tested in audits at King's College and Imperial College London. Any necessary amendments will be made before the official release on 1st October 2008. The toolkit will be promoted thereafter in collaboration with the Digital Curation Centre[3] and DigitalPreservationEurope (DPE)[4]. Training events are planned to assist organisations to adopt and implement the Framework. The audit toolkit will be freely available to use online or download from http://www.data-audit.eu Support will also be available through the website.

The methodology and toolkit will be tested further in four JISC-funded implementation projects at University College London, King's College London, Imperial College and the University of Edinburgh. These projects should conduct some twenty audits across a range of HE departments and schools and should finish in December 2008.

The DAF Methodology

The development of the DAF methodology drew on the experiences gained by staff at HATII when developing DRAMBORA,[5] a methodology for assessing the risks associated with digital repositories. At the outset the team recognised the value of a practice-oriented and intuitively applicable approach. DAF provides institutions with a straightforward method of collecting information on their research data assets. It has been designed so that it can be applied without dedicated or specialist staff and with limited investment of time or effort. The methodology has four stages:

1. Planning the audit;
2. Identifying and classifying data assets;
3. Assessing the management of data assets; and,
4. Reporting results and making recommendations.

The stages generate two key outputs: an inventory of data assets created during Stage 2; and a final report that incorporates recommendations on how data management could be improved. A detailed workflow of tasks and outputs within each of these stages can be seen overleaf (see Figure 1).

Audit stages

Planning the audit

There are two key objectives of the planning stage: (1) to secure organisational buy-in by establishing a robust business case; and, (2) to prepare as much as possible in advance of the audit so time spent on-site can be optimised. Securing agreement from top management and ensuring this commitment is filtered down is crucial. Establishing expected outcomes will assist data auditors with determining the scope and focus of the audit. By conducting background research the auditor can minimise demands placed on data creators, managers and users, and scheduling interview times and locations in advance will help ensure they are ready to contribute.

Planning of the audit involves the following tasks:

- Appoint an auditor;
- Establish a business case;
- Conduct initial research to plan the audit; and,
- Set up the audit.

Our test audits indicate that this work takes between 2-4 days, depending on the level of prior knowledge the auditor has of the department being audited and the size of the department. Where the toolkit is used internally for self-audit the initial research stages are not likely to require as much effort. The planning stage may take place over a few weeks as the auditor waits on information and responses from staff with whom interviews have been requested. During this stage a form is completed to support the capture of high level information about the organisation being audited (see DAF Methodology, Audit Form 1).

Identifying and classifying data assets

The purpose of the second stage is to establish what data assets exist and classify them according to their value to the organisation. Essentially, an inventory of data assets is compiled through a mapping exercise. The overall quality of the entire audit depends on this first knowledge-gathering exercise. Classification schemas are suggested in the inventory but will need to be tailored to the particular organisational context. The classification step will determine the scope of further audit activities, as only the vital or significant assets will be assessed in greater detail.

This stage should proceed through the following steps:

- Analyse documentary sources;
- Conduct questionnaire and/or interviews;
- Prepare data asset inventory; and,
- Approve and finalise asset classification.

Using the timing data accumulated during the test audits we can project that this work will take between 4-6 days, depending on the size and type of the organisation being audited and its data holdings. If interviews have been planned in advance during Stage 1, elapsed time should only be a couple of weeks, however this could increase if staff are unavailable to participate. During this stage an inventory of data assets, divided into groups according to their value for the organisation will be produced (see DAF Methodology, Audit Form 2)

[3] http://www.dcc.ac.uk
[4] http://www.digitalpreservationeurope.eu
[5] DRAMBORA: Digital Repository Audit Method Based on Risk Assessment is available at: http://www.repositoryaudit.eu/

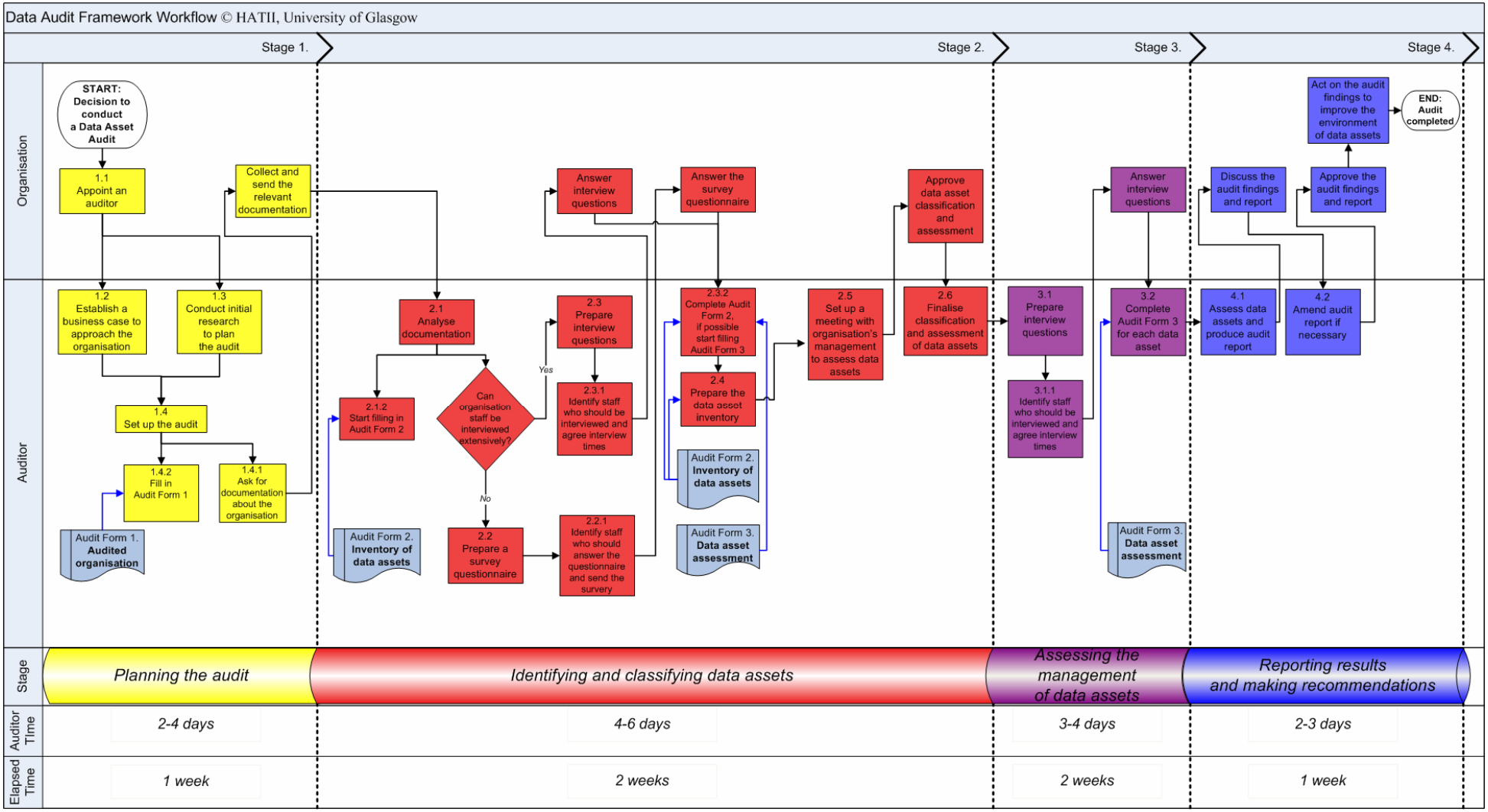

Figure 1: The Data Audit Framework Workflow

Assessing the management of data assets
The aim of this stage is to collect additional information about the data assets central to the work of the organisation. Assessing the management of these assets enables auditors assess whether the current level of resources provided is sufficient. Information collected should help identify weaknesses in data management practices and point to occasions when data are being placed at risk. During this stage several forms are completed which assist auditors in asset and context profiling (Audit Form 3A or 3B). The methodology provides two element sets to support the collection of information at different levels of detail. The level of detail adopted will be determined by the audit aims and scope set at the planning stage. Based on the pilot audits we can project that this work will take between 3-4 days, depending the number and nature of vital assets. Elapsed time is expected to be in the region of 2 weeks.

Reporting results and making recommendations
In the final stage the auditor draws together the results of the data audit to produce a final report. This report will include recommended actions to improve data management. Suggestions of relevant services and tools that could be used by the organisation to enhance their practices and services are provided in the audit toolkit and as new ones emerge we will hope to link these to the toolkit. We recommend that it would be best practice to submit the audit report to the appropriate managers within the organisation for comments before it is finalized. This stage is likely to take between 2-3 days. Elapsed time may be up to 1 week depending on the time taken to convene a meeting with management to approve the report.

Testing and updating the audit methodology

The methodology was initially tested in pilot audits based at three of the development project's partner institutions. These were split across subjects: archaeology at the University of Glasgow, engineering at the University of Bath, and GeoSciences at the University of Edinburgh. Although the audits took place in departments / schools of varying size with different data collections, the lessons learned from the pilot applications of the methodology were consistent, suggesting it is generic enough to suit diverse contexts. Moreover approaches to data curation that were encountered were consistent and confirmed the belief that auditing data assets would be of widespread benefit. We learned much from these audits and have revised the methodology as a result. We will continue to refine it as we receive further feedback from other individuals and organisations who apply it.

GUARD at the University of Glasgow

The pilot audit at Glasgow was conducted in Glasgow University Archaeological Research Division (GUARD), the archaeological research unit within the Department of Archaeology. The Unit was founded in 1989 and currently has thirty-three members of staff. It is a commercial arm of the Department and offers a wide range of archaeological services from consultation to fieldwork and post-excavation analysis. Staff are constantly engaged in projects that result in digital data assets, such as digital images, computer aided designs, GPS/GIS, and stratigraphy and finds databases.

Implementing the methodology was straightforward. The Director of GUARD was already aware of data issues within the Unit and was keen to take part. Access was granted to the shared drives on which most data was held so 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 poor. 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. Interviews were arranged with around a quarter of the workforce. Some interviews were general discussions on data curation practices but most focused on discussion of specific data assets and were crucial in completing the assessment stage. The interviews were very useful for seeing how the Unit 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 within GUARD. These aspects helped feed into recommendations we could make as to how data management could be improved.

IdMRC at the University of Bath

The pilot audit at Bath was held in the Innovative Design and Manufacturing Research Centre (IdMRC). IdMRC is a research group within the Department of Mechanical Engineering. It was set up in October 2001 with funding from the Engineering and Physical Sciences Research Council's (EPSRC) 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. It has emerging strengths in shoe and electronics manufacture.

No major issues were encountered when applying the Data Audit Framework in this context. An initial phone interview was held with the Director of the IdMRC to establish the scope, purpose and requirements for the audit. Preliminary research was then conducted using the Centre's website and at this stage a decision was taken as to how to compile the inventory. A snowball sampling technique was chosen, starting with interviews with the lead researchers of the four research themes. In all, ten face-to-face interviews were conducted. The interviews consisted of browsing personal and shared drives to identify assets, recording data sets in the inventory along with any additional information that could be easily captured, and discussing how the interviewee managed

the data. The resulting inventory listed 63 data sets, of which 18 were ranked as vital, 15 as important and 30 as minor. The inventory was not comprehensive but was representative of the data assets held by the Centre. Of the data assets described in the inventory, 30 were chosen for further analysis in DAF Stage 3. Much of the information required for this stage had already been collected, so there were only a few gaps and these were filled by soliciting information through e-mail queries.

GeoSciences at the University of Edinburgh

The pilot audit at Edinburgh was held in the School of GeoSciences, a leading international research centre rated 5/5* in the last Research Assessment Exercise (2001). The School hosts over 80 academics, 70 research fellows and 130 PhD students and attracts annual research grant and contract income of around £4-6 million. The School's 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 also be involved in inter-University Research Consortia and Research Centres.

Despite the School being much larger than the other two organisations in which the methodology was applied it was still found to be appropriate. The audit began with desk research: browsing the School website, collecting annual reports and published articles, and compiling a list of research active staff including details of their research responsibilities. Interviews were conducted with thirty-five academic/research staff to compile the inventory. The interviews were semi-structured discussions during which a broad range of additional information was collected. Although this was not a comprehensive survey, the fact that the later interviews provide information duplicating that already collected indicated to the auditor that the most significant data assets had been recorded. Of the twenty-five data assets recorded only four were classified by the interviewees as vital. A detailed analysis of these assets was carried out. The audit provided crucial evidence as to the weaknesses of current approaches employed by the School to manage its data assets. The results of the audit were drawn together and a final report was produced which recommended actions for change.

Lessons learned

Several threads were raised consistently in the feedback from the pilot audits. These are categorised into five domains.

1. Ensure timing is appropriate – The initial audits were scheduled to take place in May. When planning and setting up the audits difficulties were often encountered obtaining convenient times to meet with staff. Summer holidays, exam board meetings, conferences and extended periods of fieldwork meant that the audits commenced later than anticipated. The timing of the audit should ideally coincide with the organisation's quieter period.

Originally the time suggestions given in the methodology had been in terms of person hours. As a result of their experiences applying the methodology the auditors recommended a differentiation be made between person hours and elapsed time as the lag-time between requesting information and conducting work could be quite significant. The person hours allocated for the audit were increased from 1-2 weeks to 2-3 weeks in light of the pilot audits and a suggestion was made to allow 2 months of elapsed time.

2. Plan well in advance – Setting up interviews and waiting on documentation from the organisation can take a number of weeks. To mitigate against this and avoid the audit schedule going off track, the planning stage should be started as early as possible. The person hour requirements are minimal in comparison with the likely elapsed time so planning could run concurrently with other work commitments.

3. Adopt a method suited to the context – The decision to use interviews or questionnaires will depend largely on the culture of the organisation. Where staff are known to be responsive to questionnaires, it would be worthwhile preparing and circulating one as part of the planning stage. How best to communicate with staff also depends upon organisation context and practice. One auditor found phone calls and face-to-face meetings a more effective way to engage senior management while another found personal introductions and internal advocacy a more successful approach to communicating information about the audit than email announcements.

4. Scope the work carefully – The granularity at which assets are recorded will depend on the type and quantity of data being created. The granularity could vary within the audit due to differences in types of research being conducted. Where small sets of data are created it may be most appropriate to record assets on a project or collection basis rather than individually. Convening a meeting with key stakeholders at the start of the audit to determine the scope, purpose and requirements will help focus work. The scope could be amended during the audit if necessary.

5. Collect additional information early on – Initially the audit methodology consisted of five stages, with identifying and classifying records being separate steps. All the initial audits, however, found the optimal workflow was to collect information for these stages at once. As such the original stage two and three were merged. Auditors also found it worthwhile collecting other information early in the process. Additional information was often captured when creating the inventory, for example details of file formats, software requirements, creation dates, provenance, related data assets, storage and data management. In light of these findings we have planned that the online tool will allow Audit Form 3 to be viewed when completing the inventory (Audit Form 2) so additional details can easily be entered into the relevant fields at the time of capture.

Developing the online toolkit

Background

At the time of writing the online toolkit was still under development. We have completed the system requirements stage and this has been validated.[6] The descriptions here reflect anticipated functionality. Any discrepancies between what is planned and delivered will be noted during the tool demonstration at the iPres Conference (September 2008) and will be documented in subsequent publications about the toolkit.

Feedback from the pilots audits outlined above greatly assisted the definition of the DAF system requirements. A list of basic requirements was compiled at an update meeting and posted on the project wiki to allow additional comments to be fed back to the development team. Regular communications between the system architect responsible for defining the system requirements and authors of the methodology (one of whom had also conducted a pilot audit) ensured the appropriateness of the requirements defined.

As the online toolkit has been modelled to reflect the intentions and features of the methodology, it will facilitate planning, documentation, collection of data and final reporting. Checklists are provided and the end of each stage and contextual help will be added throughout to clarifiy what information is required. The main instance of the tool will be accessible over the internet at http://www.data-audit.eu and will be supported by secure online registries. Because we recognise that some organisations will find it unacceptable to use registries based at a second institution to store vital data about their digital assets a downloadable version will also be made available for organisations to host privately.

Functionality by audit stage

In the planning stage auditors will be guided to collect the basic information on the organisation being audited that is necessary to complete Audit Form 1. A name will be given to the audit and an upload facility will be provided for the business case. Contact details for staff within the department can be recorded and any meetings scheduled can be entered into the calendar.

In Stage 2 the auditor(s) will decide on a classification schema and set categories appropriate to the context. If a survey can be conducted the toolkit will help compile and circulate questionnaires. Alternatively the calendar system can be used to schedule interviews. Data collected at this stage will be able to be input directly into Audit Form 2. It will also be possible to enter additional data collected into Audit Form 3 ready for the next stage.

The two options for element sets in Stage 3 will be contained within separate tabs. It will be possible for the auditor to flick between one tab and another to compare the sets and make a selection as to which is most appropriate to use. Some information may already have been entered in the audit forms or pulled through from earlier stages. An additional field on both element sets will make it possible to track records by means of an automatically generated system.

The final stage of the audit requires the auditor to write a report with recommendations. Summary information and statistics will be drawn automatically from the data collected during the audit to help the auditor compile this report. The toolkit will collate information and generate a PDF appendix that contains summary details of data holdings, list of interviewees / survey respondents, and dates for the various stages of the audit. There will also be a file upload option through which the auditor may add the final audit report. It will also be possible via Stage 4 to publish audit details in the central registry. While we recognised that some organisations will not wish to have details of their data assets available in a UK-wide registry others will recognise the value of such a database to ensuring that UK higher education institutions participate in the expansion of the national knowledge economy

A status bar and calendar will be accessible throughout the audit to track progress and alert auditors to upcoming events. The toolkit will also allow files containing reports or information which helps the auditor to document the organisation, the data assets, or associated research to be uploaded. It provides ‘post-it’ style notes for comments to act as aide-memoirs for auditors. Each time an edit is made a new row will be added to the history table, making it possible to rollback to a previous version if necessary.

The design and implementation of the online toolkit will benefit from the experiences HATII gained constructing DRAMBORA Interactive, which was released in January 2008. The Data Audit Framework will be available to use online and the website will provide a shared area where users of the tool can seek advice and share knowledge gained from their experiences. DAF Interactive will incorporate a central audit registry into which institutions and departments will be encouraged to deposit their audit data so it can be federated at institutional and national level to assist strategy makers plan future work and to enable the HE community to improve its contribution to the UK digital economy.

6 Aitken, B. 2008. *The Data Audit Framework Tool: High-Level System Requirements*

Future work

The Data Audit Framework is part of a larger suite of JISC-funded data projects.[7] The development team continues to share information and lessons learned with related projects such as the four DAF implementation studies, the UK Research Data Strategy and DataShare[8]. DAF partners are committed to collaborating across project, domain and institutional boundaries to develop tools that support data creation and management.

The methodology and online toolkit enable institutions to identify their data assets and take steps to improve data management and reuse. HATII intends to seek funding to enable it to build on the audit tool to provide additional services in the future such as a data quality assessment methodology and toolkit and a tool for assessing the 'value' of data assets. Training courses for potential auditors are being developed. Information on these and additional sources of support for institutions hoping to use the Data Audit Framework to audit their research data holdings will be provided at iPres 2008 (London) and online at http://www.data-audit.eu

Acknowledgements

Development of the Data Audit Framework is funded by the Joint Information Systems Committee (JISC) through a grant from its Repositories Programme. The collaboration was made possible through the Digital Curation Centre which has acted as an umbrella for this work. The authors are grateful to partners at the Universities of Bath, Edinburgh, and King's College London for piloting the DAF methodology and providing detailed feedback on its applicability. We are particularly grateful to Dr Cuna Ekmekcioglu of the University of Edinburgh and Alex Ball at UKOLN (University of Bath). We wish to thank our colleagues Brian Aitken and Matthew Barr at HATII (University of Glasgow) for developing an online version of the toolkit: Aitken for specifying the functional requirements and Barr for implementing DAF Interactive.

[7] Details of JISC's data projects are at: www.jisc.ac.uk/home/whatwedo/themes/information_environment/researchdata.aspx

[8] For details of the UKRDS see: http://www.ukrds.ac.uk/ and for DataShare see: http://www.disc-uk.org/datashare.html

References

Aitken, B. 2008. *The Data Audit Framework Tool: High-Level System Requirements*

Jones, S., Ross, S., and Ruusalepp, R. 2008. *Data Audit Framework Methodology*, http://www.data-audit.eu/DAF_methodology.pdf

Lyon, L. 2007. *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf

McHugh, A., Ross, S., Ruusalepp, R., and Hofman, H. 2007. *The Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)*, http://www.repositoryaudit.eu ISBN: 978-1-906242-00-8

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