Data Audit Framework lessons learned report: GUARD audit

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.

Summary of progress by stage

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.

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.

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.

Final reporting and 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.

Cuna Ekmekcioglu, 31 August 2008

Data Audit Framework: Edinburgh Case study

Background and context

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.

The Audit

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.

Data Audit Framework Lessons Learned Report: IdMRC, Bath University

Alex Ball, UKOLN

August 4, 2008

1 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.

2 Performing the audit

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 Form 2. 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.

3 Lessons learned

3.1 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 organization. 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.

3.2 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.

Lessons learned report: King's College London

Stephen Grace, Centre for e-Research

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.

Summary of progress by stage

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.

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.

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.

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.

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.