Use of the Data Audit Framework within the Sudamih Project

Supporting Data Management Infrastructure in the Humanities (Sudamih)

sudamih.oucs.ox.ac.uk

Author Dr. Meriel Patrick
Affiliation Oxford University Computing Services

JISC

JISC

Project Document Cover Sheet

Project Information			
Project Acronym	SUDAMIH		
Project Title	Supporting Data Management Infrastructure in the Humanities		
Start Date	01/10/2009	**End Date**	31/03/2011
Lead Institution	University of Oxford		
Project Director	Paul Jeffreys		
Project Manager & contact details	James A. J. Wilson (james.wilson@oucs.ox.ac.uk)		
Partner Institutions	n/a		
Project Web URL	http://sudamih.oucs.ox.ac.uk/		
Programme Name (and number)	Research Data Management Infrastructure		
Programme Manager	Simon Hodson		

Document Name			
Document Title	Use of the Data Audit Framework within the Sudamih Project		
Reporting Period	*n/a*		
Author(s) & project role	Meriel Patrick (Project Analyst)		
Date	14/7/2010	**Filename**	Use of the DAF v1.0.doc
URL	http://sudamih.oucs.ox.ac.uk/documents.xml		
Access	☐ Project and JISC internal		✓ General dissemination

Document History		
Version	**Date**	**Comments**
1.0	14/7/2010	

Use of the Data Audit Framework within the Sudamih Project

Meriel Patrick 14th July, 2010

Contents

1 Executive summary 4
2 Purpose of the DAF 4
3 Use of the DAF within the Sudamih Project 4
4 Responses 5
4.1 Details of respondents 5
4.2 Nature of data assets described 5
4.3 Date of creation 6
4.4 Purpose of data assets 6
4.5 Availability of resources 6
4.6 Connection of data to published works 6
4.7 Metadata and documentation 7
4.8 Storage, back up, and maintenance 7
4.9 Importance of data 7
4.10 Funding and costs 8
5 Observations and comments 9
6 Usefulness of the DAF in this context 10
6.1 General points 10
6.2 Specific observations 10
7 Implications for Sudamih Project outputs 11
7.1 Database as a Service (DaaS) 11
7.2 Data management training 11
8 Conclusions 12
Appendix 1 – DAF-based questionnaire 13

1 Executive summary

As part of the Sudamih Project requirements gathering process, selected humanities researchers were asked to provide details of data assets they had created, using a questionnaire based on stage three of the Data Audit (or Asset) Framework methodology. Seven responses were received.

The assets surveyed were all databases which had been compiled chiefly by a single researcher, to analyse textual or historical data. Most were still being updated at least occasionally. The sample was divided reasonably evenly between data assets which were already public or which the creators intended to make public, and those which seem likely to remain private.

In most cases, the major cost of producing the assets was researcher time. The questionnaire answers suggest that some researchers may be inclined to undervalue their data assets, both in terms of actual cost of production and potential for reuse. Through the proposed database service and training modules, the Sudamih Project may have a role to play in promoting data sharing, by providing a means of doing this, and raising researchers' awareness of the reasons to do so.

Although this exercise was not a full data audit, the DAF nevertheless proved a useful tool in this context.

2 Purpose of the DAF

The Data Audit (or Asset) Framework is a tool designed to aid higher education institutions to construct an inventory of research data assets, and to describe and assess how these are being stored, managed, and shared, with a view to ensuring that research data is preserved and remains accessible. Further information is available from the DAF website: http://www.data-audit.eu/.

The full data audit process described in the DAF Methodology document (http://www.data-audit.eu/methodology.html) is intended to provide as comprehensive and systematic a review of a university department as is possible. It involves four key stages: planning the audit; identifying and classifying assets; assessing management of data assets; and reporting and recommendations. The framework is designed to be adaptable to the needs of a particular institution.

3 Use of the DAF within the Sudamih Project

Given the scope and focus of the Sudamih Project, it was impractical to attempt a comprehensive data audit as part of it. Instead, we opted for a targeted sampling approach. While this method could not, of course, provide an inventory of research data assets within the Humanities Division, it nevertheless produced some useful insights regarding the sort of data assets being produced, and the ways in which they are (or are not) maintained and made accessible.

Between February and May 2010, the Sudamih Project team conducted thirty-one interviews with members of the Humanities Division, spanning a wide range of faculties and roles. The chief focus of the interviews was on data management, but they also yielded a considerable amount of

information about research data assets. Full details of the findings are available in the Sudamih Project User Requirements Report.

With a view to exploring the data assets further, eleven interviewees with responsibility for the creation or maintenance of one or more data assets were selected and asked to complete a questionnaire based on the third stage of the DAF, intended to provide details of the nature and management of the assets under consideration. From the eleven researchers approached, seven responses were received (though one researcher provided details of several different data resources she had created). The sample size is thus not sufficient to warrant drawing firm conclusions about data management practices across the Division as a whole, but rather provides a series of snapshots or case studies. The full questionnaire is reproduced in Appendix 1.

4 Responses

4.1 Details of respondents

Three of the seven respondents were established (mid-career or senior) researchers; one was a recently appointed college lecturer describing a resource created during a Junior Research Fellowship, and three were doctoral students.

Although the respondents spanned a range of disciplines, classicists were disproportionately represented: in fact, four of the seven were from this faculty. While this tallies with the general impression gained during the Sudamih Project interviews that classics is a discipline in which the creation of data assets is relatively common, it should be noted that this was also true of some other disciplines (most notably history) which were not represented in the sample.

4.2 Nature of data assets described

The assets surveyed were all databases, and had been created as a means of aggregating and analysing textual and/or historical data. Specific subjects included Italian poetry, French song lyrics, Chinese notebooks, classical literature, economic profiles of Roman cities, and details of artefacts such as coins, inscriptions, and papyri. In addition to text, one resource included images, and one GIS data.

The sources used included books, articles, catalogues, archaeological reports, online resources, and prosopographical and GIS data. In the majority of cases, the data was compiled from existing sources rather than created. One project, however, produced analytical data through ‘a series of mixed computing and manual routines relating to a corpus of machine-readable texts’.

All the responses received related to databases which had been compiled mostly or entirely by one person. One researcher had the help of a research assistant, and another received technical assistance to transform the dataset into an online database. Two researchers were part of larger teams working within the same area (in one case a major project, in the other a more informal collaboration). No responses were received from respondents working on large, publicly available multi-contributor databases, although the interview phase revealed the existence of a considerable number of these, and indeed several questionnaires were sent to people working on such projects.

(The lack of response from this group of researchers seems to have resulted from a combination of chance and unfortunate timing: the questionnaires were circulated during the exam period, and a number of the researchers in question were examiners who were heavily committed at this time.)

The resources were in a variety of formats, including MS Excel and MS Access.

4.3 Date of creation

The oldest dataset in the sample was created fourteen years ago, and the newest within the last year. Regular work on two of the resources was still underway, while most of the others had seen a period of intense activity near the beginning of their lives (ranging from a few days to a few years), with occasional updates after that. Only one respondent said that the database was no longer being updated.

4.4 Purpose of data assets

All the datasets were compiled as part of personal research projects. The extent to which the researchers' work focused on the databases varied considerably, however. In two cases, the compilation of the data was the chief aim of the project; in others, it was a significant but not dominant element; and in still others, the resources had been created simply as a means to an end – to assist with the organization and analysis of material covered in a book or doctoral thesis, for example.

In some cases the purpose of the resource had evolved over its lifetime. For example, one respondent had initially compiled the data in table form while writing a book, but later obtained a grant from the John Fell Fund to make the material available to other scholars via the Web. Another had generated a dataset a number of years earlier during an AHRC-funded project, and then converted it into an online resource specifically so it could be included in the 2008 RAE.

4.5 Availability of resources

As indicated in the previous section, two of the resources were publicly available via the Web. However, while there were no restrictions on who had access to the material, neither of these resources was well publicized: one of the respondents commented that although the Web resource was mentioned in her book (though without a URL, as that was not available at time of going to press), it was 'irritatingly hard for people to find'.

The two researchers who were working as part of larger teams both ultimately intended to make their databases available online, along with other resources produced by their respective projects.

The resources created by the other three respondents were not presently available to anyone other than the creator. Two of the three, both doctoral students, indicated that the data might be attached (probably in print form) as appendices to their theses. The third, while happy to look things up in her database for other people if requested, had no current plans to make the material public.

4.6 Connection of data to published works

In four cases, the data was connected to published books and/or articles. However, it was always not clear whether the data resources were referenced in the publications in question. As mentioned in

section 4:5 above, in one case an online resource was mentioned in the book that drew on the data, but without it being possible to include a URL.

The data assets which were not currently connected to published works were those created by doctoral students: thus these were connected to significant pieces of unpublished research, and may be connected to published works in the future.

4.7 Metadata and documentation

Respondents were asked whether the data resources included metadata and documentation. The question about metadata was interpreted in different ways by different respondents. Some took this to be referring to metadata relating to the data set as a whole – name of creator, date of creation, and so forth. It was not clear how these respondents understood the distinction between metadata and documentation, and in some cases the answers to the two questions were very similar. Others understood metadata to be further details recorded about each item in the dataset – for example, author and date information for individual text extracts. Five of the seven respondents said their data assets had metadata under one of these two definitions.

Fewer assets were accompanied by documentation. Even in such a small sample as this, there was – perhaps unsurprisingly – a clear correlation between the presence of documentation and the availability of the data assets. Both the resources that were available online offered some documentation (described variously as 'information pages' and 'a methodology section') on the website. The two respondents who plan to make their data resources public when they are complete both also signalled their intention to provide documentation to accompany them. Respondents whose data assets were for their own private use did not tend to have documentation.

4.8 Storage, back up, and maintenance

Five of the seven respondents stored their data assets on their personal computers, and backed them up along with the rest of their data, in most cases to one or more external hard drives. Two databases – those which were publicly accessible – were hosted on institutional servers (one at the University of Oxford, one at the University of Reading), and so were backed up in accordance with the practices of the institution in question (though one of the respondents commented that she did not know if there was a back-up policy in place, and that she kept and backed up a separate, personal copy of the data). None of the resources was currently being maintained by anyone other than the respondents themselves, and in no case was a curation or asset manager involved.

4.9 Importance of data

Respondents' estimates of the importance of their data varied widely. Several said that the data was important to them and their research, but seemed less certain about its value to the wider scholarly community. Others were more confident that the data would be of interest to others in their field, and in answer to the question 'Is the data especially important?', one respondent replied 'Of course – all data is.' Unsurprisingly, respondents were more inclined to rate their data as important when creation of the data asset had been a major focus of the project: where the data had been compiled chiefly as a means to an end and had now served the purpose for which it was collected, respondents used phrases such as 'It's useful, but not exceptionally important' and 'I'd be sad to lose it, but the world wouldn't end'.

Somewhat paradoxically, when asked about the long-term value of their data, respondents were noticeably less tentative. Nearly all agreed that their data was at least potentially valuable long-term. A key point made by some respondents was that their data was not the sort that would go out of date, and discussion with a wider group of academics during the interview phase indicates that this seems to be a typical feature of humanities research data. A couple of respondents also suggested that their data might provide a useful starting point for someone embarking on a wider project.

4.10 Funding and costs

Creation of the two assets which are currently available online was supported by specific grants: in one case from the AHRC, and the other from the John Fell Fund. However, these grants applied to quite different phases of resource creation. In the first case, the AHRC grant (of £35,000) covered the cost of the initial project which collected and generated the data in 2000. Several years later, the researcher put the data into an online database himself, without receiving further funding to do this. In the second case, the data was compiled in the course of a Junior Research Fellowship (funded by an Oxford college), and a grant of £4770 from the John Fell Fund enabled the researcher to pay for this to be converted into an online database.

The other resources were not funded by specific grants, but were created in the course of the researchers' work (although as noted above, the proportion of each individual's work that this constituted varied from minimal to substantial), and thus might be said to have been ultimately paid for through the salary or DPhil funding of the researcher in question. One project, however, while without specific funding in the early stages, had just secured British Academy funding for further development.

Only one of the data assets currently has costs associated with preservation or maintenance (though this respondent did not provide any further details of these). Two other respondents noted that while there were no costs at present, there might be in the future (to cover, for example, technical updates, or if the organization hosting the data began to charge for this service), and a third said that she would like to be able to pay for technical problems to be fixed, or to expand the database, but she does not have the funding to do this at present. As one might expect, actual or anticipated maintenance costs were associated with data assets which are or will be made public, rather than those which are simply stored on a personal computer for private use.

When asked to estimate the cost of creating the data assets, most declined to specify a precise figure. Where a grant had been given, it was relatively straightforward to cite this as part of the cost, but as the two grant-funded projects mentioned above both involved a significant amount of work which was not covered by the grant, this does not tell the whole story. Equally, when a project ultimately results in (for example) one or more data assets plus a book or thesis, it is hard to say what proportion of the total cost should be attributed to the data asset.

Several respondents noted that the chief cost of the project had been their own time, but were generally wary of putting a specific price tag on this. One respondent suggested that as work on her database had been outside her normal working hours, the resource was 'essentially a cost-free by-product' of her research.

One DPhil student, however, suggested that as the creation of a database (along with a thesis which will draw heavily on it) is the major goal of his doctorate, it would not be unreasonable to regard the cost of creation of the data asset as being the cost of his DPhil, which he estimated as being in the region of £40,000, or £60,000 including fees.

None of the respondents mentioned indirect costs, such as the cost to their institution of supplying Web hosting or library resources.

5 Observations and comments

- Despite the small sample size, the questionnaire answers do provide some insight into both the sorts of data resources that are created in the humanities, and the experiences of the academics creating them. However, it should be borne in mind that there are numerous other types of asset which were not covered by this sample, such as the major multi-contributor databases already mentioned, and a range of non-database resources such as collections of audio recordings.

- The questionnaire answers and the interview phase of the Sudamih Project both indicate that data resources are created by researchers working in a wide range of disciplines and for a wide variety of purposes. It also seems true, however, that the distribution of data resource creation across faculties is not uniform: there are some areas in which the creation of data resources is more common, and consequently it may be appropriate for additional support to be provided for some faculties.

- In a significant proportion of cases, humanities data is compiled from existing sources rather than generated expressly for the project. Those sources are often publicly accessible via libraries and the Web, meaning that the chief cost in humanities data projects is frequently researcher time. None of the projects surveyed required specialist equipment, other than access to a computer with suitable software. (In the wider Sudamih project interviews, however, we learnt of projects which had incurred significant travel costs to gather data overseas, and it should perhaps be borne in mind that library access carries a cost to the institution.)

- Humanities data assets are frequently almost infinitely expandable: in many cases there is no point at which they may truly be said to be complete. On the other hand, humanities data is often of a sort that does not become out of date. Both these facts mean that there is significant potential for reuse of data assets: a database created by one researcher may form a valuable starting point for another embarking on a project in the same area.

- Some researchers have already made their data assets publicly available, or intend to do so. However, the questionnaire answers indicate the existence of a significant number of resources which are likely to remain accessible only to the creator (or perhaps be made available to a very limited extent, such as in the appendix of a thesis). In some cases, this results from a deliberate decision (some of the reasons for such decisions are detailed in Section 4.7 of the Sudamih Project User Requirements Report), but in others the possibility

of making the data public may simply not have occurred to the researcher, or the lack of a straightforward means of doing so may prove a bar.

- Even in cases where the data has been made available, the responses indicate that its existence may not be widely known.

- The responses to the question about the cost of producing the data assets suggest that in some cases researchers may be undervaluing their own time, and by extension, the resources they produce. It is possible that this in turn results in their being less likely to take steps to make those resources public.

6 Usefulness of the DAF in this context

Ultimately, the DAF proved a useful tool for supplying additional information about data assets created by the researchers we spoke to, and thus fleshing out the picture of their working practices and what can best be done to support them and to maximize the value of their work to the academic community. It was not, however, as informative as the main Sudamih Project interview phase, but given the aims of the project and the nature of the DAF, this is neither particularly surprising nor a criticism of the DAF methodology.

6.1 General points

- The biggest challenge faced in any data auditing procedure is simply that of eliciting the relevant information from researchers with many other calls on their time. The DAF implementation guide offers advice on this (including thorough planning, securing the cooperation of senior management, and limiting the scope of the audit), but it is plain that any attempt to produce a comprehensive survey would be a major undertaking, and there is therefore a need for careful considerations of the aims of an audit and how these might best be achieved.

- A major advantage of the DAF is its flexibility: the methodology is designed to be easily adapted to a range of circumstances, and can be tailored to suit the specific purposes of a given audit. In particular, the use of the DAF methodology in the Sudamih Project was for a purpose somewhat different from that for which it was originally designed, but its flexibility meant it was nevertheless possible to gain useful results with it.

6.2 Specific observations

- The variety of types of response to the questions about metadata and documentation suggest that these terms are not universally familiar to academics. To obtain useful information, it is therefore important to ensure that technical terms such as these are clarified. This is particularly essential when a data audit is being conducted via written questionnaires, as this does not afford the same opportunity for additional explanation as an interview-based method.

- Some respondents apparently found the question about the cost of creating the data assets hard to answer. In addition to the difficulties in estimating the financial value of researcher

time, none of the respondents in this sample made any reference to costs to the institution (for server space, library access, and so forth), despite costs borne by the university or department being specifically mentioned in the question. This suggests that asking researchers alone may not always give a complete picture of the costs of creating data resources. Nevertheless, this question was illuminating regarding researchers' own views of their data resources.

7 Implications for Sudamih Project outputs

7.1 Database as a Service (DaaS)

The questionnaire answers provide a series of case studies of database use within humanities research. In broad terms, the picture painted by these corresponds with the user requirements for a database service which emerged from the interview phase (detailed in section 5 of the User Requirements Report). Key points include:

- Textual data predominates, but some projects also include images and GIS data
- Support for non-standard character sets (e.g. Chinese or Greek) is vital for many researchers

In section 5 above, it was noted that some researchers have datasets which they are not currently making public, and that even when data resources are made available online, these are not always well publicized. A central database service might meet the needs of those researchers who currently lack the means to make their data public, or who would like to make data which is already public known to a wider audience. Expected requirements of this group might include:

- Ability to import existing data in a range of formats
- A straightforward means of connecting the database to a Web interface
- A low cost, sustainable service
- A search/browse facility which allows other users to find out about data assets – and ideally, indexing of the data assets elsewhere (perhaps in library or electronic resource catalogues)

7.2 Data management training

Even if a straightforward, inexpensive means of publishing and publicizing data can be made available to researchers, this is unlikely to have a significant effect on general practice unless researchers see the value – to themselves or others – of making use of such a service. If, as the questionnaire results suggest, it is true that a significant number of researchers are undervaluing the data assets they produce (despite acknowledging that in theory the material might be of use to other researchers), it seems probable that this will make them less likely to take active steps to make their data available.

There may therefore be some value in providing training which raises researchers' awareness of both the true cost (including researcher time and institutional expenses) and the value (to the academic community) of data assets, with a view to encouraging sharing of data wherever this is practical.

Additionally, the apparent confusion over the meaning of ‘metadata’ and ‘documentation’ suggests that some researchers might benefit from training covering these components of data assets.

8 Conclusions

The questionnaire based on the third stage of the DAF proved a reasonably useful tool for supplying further detail about researchers’ data assets. The exercise also demonstrated the adaptability of the DAF methodology.

The questionnaire responses indicate the existence of humanities data assets with potential value for reuse by other researchers. Some assets are or will be public, and in some cases researchers have reasons for keeping the data private. In others, however, data appears to be remaining inaccessible because researchers lack the means or the motivation to make it public. There is also evidence that some researchers tend to undervalue their datasets.

Through the DaaS and the planned training modules, the Sudamih Project may be able to promote increased data sharing, by both providing a straightforward means of doing this, and raising researchers’ awareness of the benefits of doing so.

Appendix 1 – DAF-based questionnaire

Thank you for your participation in the SUDAMIH Project. You mentioned that you've been involved in the creation of a fairly substantial data resource, and we would greatly appreciate it if you were willing to complete the questionnaire below. This will give us a better picture of the kinds of data being generated in the humanities at Oxford.

If any of the questions are unclear, inapplicable, or otherwise hard to answer, please let us know. Please return the completed questionnaire (with any queries) to meriel.patrick@oucs.ox.ac.uk

Many thanks for your assistance.

- What kinds of data do you generate?
- For what is this data used?
- What is the title of the dataset?
- Who exactly created it and when?
- Why was it created?
- Who was it created for?
- Was the development funded by a particular funding body?
- What sources were used?
- Where is the data stored at the moment?
- Is it actively maintained by anyone?
    - If so, who?
- Is there a curation manager or asset manager involved?
- How do you go about finding things again if you need to refer back to the data?
- Is there any 'metadata' attached to the data resource, e.g. keywords, information about the authors, date information?
- Do people still refer to the data?
- Does the data ever get updated?
- Are there restrictions regarding who can access the data?
- Is the data connected to any published work?
- Is there a back-up or archiving policy?
- Is the data especially important?
- Does the data potentially have long-term value?
- Is there any means for people to find out about the data?
- Is there any documentation to go with the data?
- Are there any preservation or maintenance costs associated with the data?
- Roughly how much would you estimate the costs of creating the dataset to have been (bearing in mind the salaries of the people involved, and equipment required, and additional costs borne by the university or department)?