Sudamih Benefits Case Study

Supporting Data Management Infrastructure in the Humanities (Sudamih)

sudamih.oucs.ox.ac.uk

Author	Dr. James A. J. Wilson
Affiliation	Oxford University Computing Services

RESEARCH DATA MANAGEMENT BENEFITS CASE STUDY - SUDAMIH

SUDAMIH CASE STUDY : SUPPORTING DATA MANAGEMENT INFRASTRUCTURE FOR THE HUMANITIES

1.   Background

The Supporting Data Management Infrastructure for the Humanities Project (Sudamih) is working on two particular aspects of research data management infrastructure: training for researchers; and the development of simple and intuitive 'Databases as a Service' (DaaS) software which researchers can use to create, edit, and share their own research databases. The project is working with researchers from the Humanities Division at the University of Oxford to meet the particular needs of humanities scholars in the first instance. The University intends ultimately to expand the infrastructure developed to the other academic disciplines as well.

Sudamih is a sister project to EIDCSR (Embedding Institutional Data Curation Services in Research), which assisted researchers involved in anatomical imaging and modelling prepare their data for long-term preservation and re-use. The Sudamih and EIDCSR projects are both part of a broad programme at the University of Oxford to improve the infrastructure for managing the University's research data.

This case study looks at both the training and DaaS aspects of the Sudamih project in terms of the benefits that they could bring to funding agencies, the University, but primarily to the researchers themselves.

The Sudamih Project commenced in October 2009 and is due to conclude at the end of March 2011. The core project team working on the Sudamih Project consists of a software developer, Asif Akram, a research analyst and training office, Meriel Patrick, and the project manager, James A. J. Wilson, all of whom are working on the project on a 0.5 FTE basis. The Project Director is Professor Paul Jeffreys, who is also the director of IT at the University

of Oxford. Co-PIs include the head of the Infrastructure Systems and Services Group at Oxford University Computing Services (Mike Fraser), the head of the Humanities Research Support Team (Andrew Fairweather-Tall) and academic researchers from the History and Classics Faculties (Ian Archer and Andrew Wilson). The Bodleian Libraries are also involved in the project.

2. ESTABLISHED PRACTICE AND CHALLENGES

2.1 TRAINING

The Sudamih Project began by interviewing humanities researchers to understand current information and data management practices. This included asking where they felt there were deficiencies in current data management training provision and what could be offered to improve to the situation. It quickly became apparent that many of the researchers we spoke to had previously given little thought to data management issues, and certainly hadn't received any training relating to the issue. 77% of the interviewees said that they had never received any data management training, and whilst the other 23% had received training in using software to help manage data, this training was not offered as data management training *per se*. Almost all of the researchers whom we interviewed thought that data management training would be of value, particularly to new research students.

Various different approaches to organising files and safeguarding data were being applied by the researchers; quite a few were not aware of any central infrastructure to help them (about a third were not aware of the University's data back-up and long-term file store); and a number of them said that it would be good if there were some sort of technical consultancy service that could help them with data-intensive research projects (which the University does in fact already offer). A handful of researchers with whom we spoke were working on large publicly-accessible web-hosted research databases, and, as one might expect, they generally had a much better grasp of data management techniques, but the majority lacked both technical awareness and a real sense of organisational principles. 78% of the researchers we interviewed admitted that they sometimes had to spend time finding

information that they knew they had but couldn't recall where they had filed it. It may also be the case that improvements in organising and structuring research information might help improve the actual quality of research: as one philosophy lecturer put it, "I do believe that our research could be enhanced by having better ways of storing information, because the way I store my thoughts makes a difference to how I use them when progressing in my thinking, so I can see that improving the way I store them might help the actual thinking – apart from saving time, it might be more substantial than that: having a clear view of what I've done, or how my different project interconnect, might just be heuristic in a sense".

As well as speaking to researchers, we also assessed existing data management support services and materials for researchers to see where there were gaps, what could be repurposed, and what needed to be further developed.[1] We discovered that whilst there were some existing training materials that touched upon the periphery of data management, and some technical data management expertise within the Computing Services and other departments, there was relatively little either at Oxford or elsewhere that could simply be picked up and re-used which would directly address the problems that humanities researchers faced. We have therefore had to develop a suite of relevant training and support materials largely from scratch. The one obviously mature and significant pre-existing data management course was the Digital Curation Centre's 'Digital Curation 101'. The DCC graciously agreed to run one of their '101 Lite' courses in Oxford and to adapt it to better meet the concerns of researchers. The speed in which the 25 places on the course were filled (within 48 hours of being opened for registration) suggested that there was demand for such training. Feedback was positive, although several attendees thought the course would benefit from more examples or case studies and wanted more emphasis on the practicalities of managing and curating data.

---

[1] This processed involved a literature search, an extensive search of the Web, and speaking with other support services at the Universtiy of Oxford.

RESEARCH DATA MANAGEMENT BENEFITS CASE STUDY - SUDAMIH

2.2   DATABASE AS A SERVICE

Although many researchers in the humanities do not work with structured data of the sort that lends itself to storage and analysis in databases, the researchers we interviewed clearly felt that there was a significant and growing trend in the humanities towards projects which collect data and which could benefit from such applications. Many humanities researchers do not have the technological awareness to know when it would be appropriate to use a database for the information they gather during their research, others are using databases or pseudo-databases (such as bibliographical software) but in eccentric and sub-optimal ways; many researchers who do work with structured data do so via desktop database software such as Microsoft Access, which is not ideal for sharing data, either with collaborators or more widely. The information held in private databases is frequently not explained or documented in such a way that it could be properly understood and interpreted by other researchers even if they could get hold of the raw data. This seemed to be partly because such documentation would require extra effort, but mostly because researchers often either didn't think about data sharing at an early stage of their research or didn't think that the data they were collecting would be worth sharing.

The DaaS is intended to improve the data organisation practices of humanities researchers whilst also facilitating the preservation and reusability of the research data they generate. The system, when complete, will enable researchers to quickly and intuitively import existing databases in common formats such as Access, or create new databases from scratch. It will have a graphical interface via which databases may be restructured, forms created, data added and edited, and searches conducted. The initial (Sudamih) development phase of the DaaS will result in a web-hosted service able to output query results in tabular, statistical, graphical (map) formats, and forms, and will support textual, image, and geospatial data. As the software will be hosted centrally and accessed via the Web, multiple collaborators can work on the databases simultaneously and it will be possible to 'open up' any given database to public searching (using a generic querying interface). The University has the responsibility for ensuring the long-term accessibility of the database, whilst consistent searchable

metadata can be gathered relating to each database within the system, making the databases easy to discover.

3. BENEFITS FROM PROJECT

3.1 TRAINING

A number of potential benefits arising from the training that Sudamih has been developing can be identified:

1. Time saved by researchers by locating and retrieving relevant research notes and information more rapidly

2. Improved quality of research by locating better, more relevant research information than would otherwise be the case

3. Improved quality of research by linking materials in such a way as to highlight connections and trigger new ideas

4. Improved comprehensibility of research information and data after long time periods

5. Better awareness of software tools to assist research management

6. Better awareness of central infrastructure services intended to help researchers, including technical help and assistance with funding bids

Assessing the long-term impact of information management training is difficult, as whilst the benefits of improved organisational systems and techniques are likely to accrue over time, there are so many factors that could influence information management practices besides a specific training course or website, the benefits derived from any given set of training materials are hard to isolate and measure. Sudamih is therefore focusing mostly on gathering what information it can about short-term impacts. One way of assessing the longer term success of the training materials developed by the project might be to see whether they are still being used (albeit in an adapted form) at Oxford and beyond in five years' time.

RESEARCH DATA MANAGEMENT BENEFITS CASE STUDY - SUDAMIH

Besides the DCC-run workshop on data management for researchers, Sudamih is staging two face-to-face three-hour courses on data management skills and tools for researchers in the humanities, and has been involved in a number of induction events for new research students and post-docs. The project has also developed a large number of web pages for a new online initiative for researchers at the University of Oxford – the Research Skills Toolkit (RSTk). As the RSTk is currently only accessible via the Oxford Single Sign-on scheme, we can gather rich data about who is using the online guidance and their online behaviour.

Feedback forms will be used to help gauge the near-term impact of the face-to-face courses. We will be asking course attendants if they have changed (or considered changing) their information organisation strategies as a result of the courses and training materials and, if they have, why they have done so and what advantages they anticipate from doing so. The second face-to-face course will be concerned particularly with software tools, so we will be able to assess whether attendants awareness of what is available has increased (and increased usefully) via feedback to that. We should be able to get a rough idea as to whether researchers are making more use of central infrastructure by comparing the rate of uptake of such services over the last few months of the Sudamih project with previous periods, although of course we will not be able to isolate the effects of the project. Once the RSTk materials go live (early February) we will be able to better assess which aspect of data management researchers are most concerned by (by looking at proportional use), and get a sense of which tools and tips they think are likely to be most useful to them.

3.2   SHARED BENEFITS DERIVED BOTH FROM TRAINING AND DaaS

7. Reduced risk of data loss (researcher benefit / institutional benefit)

8. Improved version control (researcher benefit / institutional benefit)

These benefits will be felt by both researchers and the institution. Given that the University technically 'owns' the data compiled its researchers, data loss impacts upon the assets held

by the Institution as well as having potentially very serious consequences in terms of research time effectively wasted.

We will assess these benefits by considering the costs of creating research databases in the humanities against the comparative risks of losing data using differing backing-up strategies.

3.3   DaaS

Benefits arising from the uptake of the proposed Database as a Service system are anticipated to include the following:

9. Improved sharing and re-use of data

10. Improved data security (storage; access & identity management)

11. Less duplication of data

12. Faster preparation of structured data

13. Greater technical support efficiency

14. Economies of scale due to central hosting

15. Greater data consistency between projects enabling repurposing of data and mash-ups

16. Data becomes more reliably citable due to use of DOIs

17. Strengthened research grant applications

18. Inspiration for new research

Whilst we will not be able to measure all of these benefits within the timescale of the Sudamih Project, we shall assess potential savings deriving from the use of the DaaS by gathering information relating to the costs involved in creating, sharing, and maintaining a large research database. We will do this by using the Roman Economy Project (REP) as a case study. The REP is currently working to combine a number of databases into a coherent whole with a public search interface and is testing the DaaS as a means for so doing. We

shall measure costs of developments so far, including data gathering, reorganisation, and current hosting models and contrast these with the anticipated costs of supporting the DaaS.

We should be able to get a clearer assessment of some of the other potential benefits (such as improved support efficiency and economies of scale) by talking to support teams in the Computing Services.

Several of the identified benefits derive from the improved awareness of data expected to result from the use of the DaaS. We will require that basic metadata be added to describe each database hosted on the system, and (with the creator's consent) open this metadata up so that people can find out that the database exists even if it's creator does not wish to make the data itself publicly available.

We already have evidence that there is considerable demand for the capabilities offered by the DaaS thanks to feedback from a workshop organised by Sudamih entitled 'Databases in the Humanities – Where Next?' We used the workshop as a opportunity to explain the features of the DaaS and asked attendees (who were mostly researchers in the humanities who already had an interest in research data and databases) whether they would consider using the DaaS (or an equivalent) once it was fully operational to develop or host a database. So far, 47% of respondents indicated that they would seriously consider it, whilst a further 47% said it was a possibility. Respondents gave various reasons why they thought the DaaS might interest them, including: expanding upon projects begun using personal database software such as Access; sharing data input; for hosting small-scale personal projects; and as a cheap long-term host for projects that are no longer otherwise supported.

4.   SUMMARY AND KEY POINTS

Whilst Sudamih has already gone some way to illustrating need and demand for its training and DaaS outputs, more work remains to be done to verify the anticipated benefits and,

where possible, to quantify them. It will not be possible in many instances to produce estimates of benefits in financial terms, but we should be able to give clear qualitative justifications for the proposed services.

Key points include:

- The fact that research data management materials are being created and disseminated addresses a clearly identified gap in current provision which researchers themselves recognize as important

- Researchers involved with structured data recognize that the DaaS offers the potential to assist them in several respects and will certainly consider adopting it once key functionality is in place