



Sudamih Researcher Requirements Report

Supporting Data Management Infrastructure in the Humanities (Sudamih)

sudamih.oucs.ox.ac.uk

Authors

Dr. James A. J. Wilson
Dr. Meriel Patrick

Affiliation

Oxford University Computing Services

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	Sudamih Researcher Requirements Report		
Reporting Period	n/a		
Author(s) & project role	James A. J. Wilson (Project Manager); Meriel Patrick (Project Analyst)		
Date	26/7/2010	Filename	Sudamih Researcher Requirements Report .docx
URL	http://sudamih.oucs.ox.ac.uk/documents.xml		
Access	<input type="checkbox"/> Project and JISC internal	<input checked="" type="checkbox"/> General dissemination	

Document History		
Version	Date	Comments
1.0	26/7/2010	Report signed off by Sudamih Project Working Group and Steering Group
1.1	3/8/2010	Fixed minor typos and added missing interviewee.

Sudamih Researcher Requirements Report

James A. J. Wilson & Meriel Patrick

26th July, 2010

Contents

1. Executive Summary.....	5
1.1. Findings	5
1.2. Recommendations	5
1.2.1. Database as a Service.....	5
1.2.2. Data management training	6
2. Introduction	7
3. Methodology.....	7
4. Current Practices.....	9
4.1. The nature of humanities research data.....	9
4.2. Note taking.....	10
4.3. Organizing data	11
4.3.1. Organizing digital information	11
4.3.2. Organizing emails.....	12
4.3.3. Organizing paper-based information.....	13
4.4. Finding data when it's needed.....	13
4.5. Bibliographic software	15
4.6. Structured data	15
4.7. Data re-use.....	17
4.8. Storage and backing-up	19
4.9. Data ownership and rights.....	19
4.10. Data Policies.....	20
4.11. Conclusions	23
5. Database as a Service (DaaS) Requirements.....	23
5.1. Key database requirements	24
5.2. Editing rights and security.....	25
5.3. Data dissemination	27
5.4. Support requirements.....	27
5.5. Additional features and concerns.....	28

5.6.	Conclusions	29
6.	Training Requirements.....	30
6.1.	Existing training provision.....	30
6.2.	What do people understand by the term ‘data management training’?	32
6.3.	Have people already undertaken any data management training?	33
6.4.	Is there a need for data management training in the humanities?.....	35
6.5.	What should training cover?.....	37
6.5.1.	Information management training	38
6.5.2.	Technical training on database design and software tools.....	39
6.5.3.	Technical advisory service for Principal Investigators	40
6.5.4.	Training specifically for graduate research students	41
6.6.	How should training be implemented?	42
6.6.1.	Institutional context of training	42
6.6.2.	Medium of training	43
6.6.3.	Timing of training.....	44
6.7.	Conclusions	45
	Appendix A – Interview Question Template	47
	Appendix B – Index of Interviewees	49

1. Executive Summary

This report sets out current research data management practices in the humanities at the University of Oxford and makes recommendations regarding researcher requirements for a 'Database as a Service' (DaaS) system and for data management training.

1.1. Findings

- Humanities research is hugely diverse, and makes use of an enormous range of types of sources.
- The intellectual value of humanities datasets tends not to depreciate over time – a database of Roman cities is potentially of as much use to researchers in fifty years time as it is today, provided it is not rendered obsolescent through technological change. Humanities scholarship often aggregates to a 'life's work' body of research, with any given researcher often wishing to go back to old datasets in order to find new information.
- There appears to be a growing trend towards structuring data within the humanities. This is partly because technological advances make more sophisticated data projects possible, and perhaps partly because of the changing priorities of funding bodies.
- Methods of organizing data also vary considerably, as does the extent to which researchers succeed in creating and maintaining a well-functioning system.
- Good information management is time consuming, and academics often find themselves with insufficient time to keep on top of it.
- Most researchers are willing in principle to share their data with others, but in practice choose not to do so, for a variety of reasons.
- Data storage is generally on personally-owned machines and backing-up is generally also to personal devices on an *ad hoc* basis. Knowledge of centrally-provided services is limited and they are seldom used.
- There is a significant amount of confusion over the ownership of research data. This is exacerbated by complex situations in which multiple people or organizations may have different claims on the same resource.
- Most researchers are unaware of formal policies relating to data management, although those who have applied for funding may know something of the requirements of a specific funding body.

1.2. Recommendations

1.2.1. Database as a Service

- The DaaS should be developed and trialled. The response to the proposed DaaS was generally positive, and while such a service will naturally not be of use to all researchers a substantial proportion could see some potential application for their own work
- The DaaS should be trialled both with researchers with existing databases, to ensure it offers the functionality they expect, and also with researchers still in the planning stages, to ensure that it is appropriately intuitive and can meet their expectations
- It should be simple for researchers to import existing databases into the DaaS, otherwise they are unlikely to move to the new system despite the other advantages it might offer
- If the DaaS is going to be broadly useful to humanities researchers, it needs to meet the following key user requirements:

- Intuitive and easy-to-use interface
 - Flexible searching and querying
 - Able to deal with a range of data types
 - Records may be linked to external sources
 - Supports diacritics and text in non-Roman alphabets
 - Can be edited by multiple people
 - Must be easy to edit online via a Web interface
 - Query outputs presented in Web browser
 - Data can be downloaded and worked on with desktop applications
 - Data integrity can be preserved
 - System is stable and secure
 - There is good user support and training
- The University must clarify the ownership rights of researchers to any data they place in the DaaS. Researchers would be unlikely to use the services if they felt their ownership of that data would be compromised

1.2.2. Data management training

- Training needs to be based upon actual research problems commonly faced, not promoted as generic skills training
- Training should on the whole be provided via face-to-face courses, with supplementary online content
- Graduate students should receive data management training early in their research careers, but ideally not before they have already had the opportunity to assess and start to gather the kinds of sources that their research will demand
- Aspects of data management training should be integrated into existing training where possible
- If possible, information management issues should be included in compulsory training
- If possible, courses should be customizable to allow for the needs of particular faculties. It should be ensured that default examples are broadly applicable across disciplines
- Training should be offered in both 'broad' data (or information) management skills and also in 'narrow' (technical) data management skills
 - 'Broad' data management would include: organizing your files in such a way as you can retrieve information quickly and easily; backing up; versioning; managing email; linking notes to content; keeping track of your sources
 - 'Narrow' data management would include: which type of software is most suited to particular requirements; structuring data in relational databases; querying and retrieving information; long-term curation – data formats, obsolescence and migration issues; using the DaaS
- An advisory service should be on hand to offer one-to-one technical advice about data management (particularly in the narrow sense) to Principal Investigators wishing to apply for project funding

2. Introduction

The Supporting Data Management Infrastructure in the Humanities (Sudamih) Project seeks to address the data management requirements of humanities researchers within the University of Oxford, particularly as they relate to two strands of activity:

1. Data management support and training
2. The provision of a centrally-provided online database creation, editing, and querying facility, or 'Database as a Service'

The need for the development of these elements of institutional infrastructure was identified by an initial scoping study carried out by the Office of the Director of IT at the University of Oxford during 2008 and 2009 (<http://www.ict.ox.ac.uk/odit/projects/digitalrepository/>). The Sudamih User Requirements Report follows up this work, investigating how data management training and a 'Database as a Service' might actually be implemented to best meet the needs of humanities researchers, from across multiple disciplines and at various stages of their research careers.

Understanding how researchers in the humanities think about data, how they conceptualize, gather, store, use, and generally look after it, has formed an essential part of the study, and in that respect this report may be of broad interest to those involved in the humanities research support, rather than merely to the Sudamih project team for whom the requirements were ostensibly gathered.

Section three of this report explains the methodology upon which it is based; section four describes our findings relating to how researchers use data; whilst sections five and six deal particularly with the requirements we identified which relate to the database as a service and data management training.

Throughout this report, numbers in square brackets refer to the interview in which the point was raised or the comment was made. An anonymized index of interviewees is provided in Appendix B – Index of Interviewees

3. Methodology

This report is based on findings from thirty-two semi-structured interviews. Each interview lasted for approximately one hour and sought to establish: the kinds of research being undertaken; the kinds of source materials being used for the research; how those materials were used, stored, recorded and annotated; how notes (both handwritten and electronic) were then managed and used; and how all this information was then marshalled and processed to create the final research outputs. We then proceeded to ask questions specifically relating to the potential usefulness of the 'Database as a Service' system, and to data management training.

The Sudamih project is taking a broad approach to 'data', regarding it as encompassing not just structured information on computers, but the whole range of materials that researchers must assemble and analyse in order to produce their research outputs. This potentially includes, therefore, printed books, articles, and even handwritten notes, in addition to electronic sources and datasets. We began each of our interviews with researchers by explaining this, to encourage them to think widely about the kinds of 'data' they used and so as not to scare away those who made little or no use of data in the 'narrow' sense of structured information stored in tables in electronic formats.

However, given our interest in finding out about researcher requirements for our proposed 'Database as a Service' system, we did tend to spend more time focussing on the use of data in this 'narrow' sense when interviewing scholars who made significant use of databases in their research.

The large majority of our interviews were conducted with scholars actively conducting research in humanities disciplines. Besides active researchers, we undertook one interview with divisional research training staff, one with members of the divisional Research Support Team, and one with a member of the University's Research Support Team. We tried to interview a broad sample of researchers, covering all of the major humanities disciplines and with representation from all stages of the academic career.

Faculty	Postgraduate	Early Career	Senior
(Classics) – Classical Languages	1	1	
(Classics) – Ancient History	1		1
(Classics) – Classical Archaeology	1	1	1
English		2	1
History			3
Linguistics, Philology, and Phonetics	1	1	
Medieval and Modern Languages			2
Music	1		2
Oriental Studies	1		3
Philosophy		1	1
Theology	1		1
Rothermere American Institute		1	
Total	7	7	15

Table 1. Number of researchers interviewed in each faculty by career stage

We did *not* attempt to interview an entirely representative sample of the humanities division in terms of expertise working with data. Instead, we intentionally contacted several researchers on the basis that their work relied heavily on databases and highly structured information, in order to ensure that we could get 'expert' opinions as to how we should approach the development of the DaaS and to see if their attitudes to data management training differed from researchers less heavily involved in data collection and analysis in the 'narrow' sense. It was also the case that we needed to speak to researchers using different types of electronic data: text, image, and geo-spatial, hence we specifically sought out such people. Although we have not surveyed the humanities division as a whole at Oxford, it is almost certainly the case that technically proficient database users are over-represented in our sample.

We selected our interviewees initially on the basis of recommendations by the Sudamih Project's Principal Investigators, and then to a lesser extent on following 'friend of a friend' recommendations. This ensured that we got to speak to a number of senior researchers with quite

different approaches to data. We identified the majority of our interviewees, however, via the World Wide Web, looking for people with diverse research interests and experiences across the spectrum of the Oxford Humanities Division.

Before beginning the interview process, we produced a set of questions that we hoped to get each interviewee to address (where relevant). We did not use this as a script, and moved freely between topics as the conversation developed, to encourage a natural flow of ideas. Although our set of questions did evolve slightly over the course of the interviews as we became more or less interested in probing particular aspects of research, we followed the same basic prompts throughout. A copy of the final version of this 'prompt sheet' is included with this report as Appendix A.

4. Current Practices

The first part of each interview focused on the nature of the research in which the interviewee was currently engaged. During this phase we sought to learn about the working practices of the researcher and in particular their procedures relating to data.

4.1. The nature of humanities research data

The first observation to make regards the diversity of humanities research data. We encountered an enormous range of material that people were working with to produce their research outputs.

Sources included:

- Scholarly articles, conference papers, and books
- Print editions of literature and music
- Images – photographs, scanned documents
- Manuscripts, including different versions of a particular text
- Newspapers and magazines
- Inscriptions on stone or other materials
- Correspondence
- Library and archive catalogues and card indexes
- Archaeological catalogues
- Self-recorded audio recordings of speech
- Published audio recordings of music
- Questionnaires
- Video recordings of television news broadcasts or musical performances
- Physical objects such as monuments and archaeological remains
- Songs with accompanying music notation in various formats
- Records of prisoners
- Undeciphered ancient texts
- Plans and drawings
- Wills
- Atlases
- Memoirs and journals
- Yearbooks
- School curricula

From these sources, researchers picked out relevant extracts, or recorded observations, or noted down particular properties of the source material with which they were working that they were interested in exploring in more depth. In some cases, transcription, digitization, and preparation of scholarly editions of primary sources was itself a significant element of researchers' work.

In many cases there then follows the traditional process of reflection, grouping, and comparison of notes and sources leading to an interpretation or reinterpretation that is then written up as an original research output in the shape of a conference paper, journal article, or book chapter. However, a greater than anticipated number of the researchers we spoke to did not move straight to this interpretive stage but rather structured their research data in some more formal way first, such as in a database, so that particular research questions could be asked of it [see section 4.6]. Our interviews suggest that this sort of data-driven research does seem to be growing in popularity in most humanities disciplines.

Given the range of sources and approaches employed by humanities researchers, it is dangerous to generalize too freely about humanities research data. One can, however, hazard some cautious observations that are true for most if not all humanities data collections:

- Data tend to be compiled from existing sources, not created from scratch (with the exception of some linguistic data gathered under 'laboratory' conditions).
- The intellectual value of humanities datasets tends not to depreciate over time – a database of Roman cities is potentially of as much use to researchers in fifty years time as it is today, provided it is not rendered obsolescent through technological change. Humanities scholarship often aggregates to a 'life's work' body of research, with any given researcher often wishing to go back to old datasets in order to find new information.
- Collections of data are frequently incomplete or inconsistent due to the inconsistent nature of the sources.
- Not all properties that potentially could be captured are so. Researchers tend to record only those properties that are likely to be relevant to their current research rather than everything that might potentially be useful to other projects in future. One senior researcher commented that "when time is tight, you have to make some tough decisions". For example, when gathering data from records of prisoners, her team opted not to collect personal names, as typing these is time consuming. While this information wasn't needed for the project then in hand, including it would have permitted other uses of the data, such as linking it up to census records. [5]
- When not originally gathered with wider dissemination in mind, humanities data is often very 'personal' to the original compiler, which can pose issues should that data be shared or re-purposed by other scholars.

4.2. Note taking

Many of the researchers we spoke to took a mixture of handwritten and electronic notes, sometimes as a result of restrictions imposed by their physical working environment, but more frequently due to personal preferences or habitual working practices.

One senior researcher said that while he writes electronically, he takes notes mostly on paper, because he reads on paper: "I find I can see more that way, and it allows me to read on the bus or

elsewhere.” [19] Another researcher took handwritten notes due to the resource type he worked with: as he worked mostly with sound recordings, which he could listen to anywhere, he would sometimes take notes by hand as he was not near a computer [22].

That said, a number of the researchers we spoke to had at some point over the last ten years or so switched to taking all their notes on their laptops and abandoning pen and paper. This was invariably consciously motivated by information management considerations – particularly the ease with which electronic notes can be searched and the risks of misplacing handwritten notes [8]. Some researchers took occasionally took notes by hand, but then typed them into their computer later [7].

Some researchers who took notes primarily on their computers still preferred to work from hardcopy versions of their sources when producing their research outputs, so would print things out and then revert to a more traditional data management system at that point. There was no observable correlation between age or stage of research career and preference for traditional methods over technological ones: for example, graduate students did not seem noticeably more likely to take notes electronically than senior researchers.

4.3. Organizing data

The manner in which researchers organize the information they gather varies considerably. Most researchers these days prefer to obtain journal articles in electronic formats such as PDFs, although longer works still tend to be purchased or borrowed from libraries as printed books. If primary sources are available electronically, and the electronic reproduction is of a high enough quality and with enough metadata to meet the requirements of the research in hand, most researchers will work with these. Electronic resources offer a number of advantages: they are searchable, and resources such as manuscripts can be consulted without requiring access to the original copy. However, not all digital resources are of a suitable quality for use in serious research. One senior researcher commented that some online texts have a limited search function, and that it’s often more efficient for him to work with a paper copy [19]. Another researcher observed that digitization within his discipline was somewhat haphazard, and that print editions will often include additional information [28].

4.3.1. Organizing digital information

Our interviewees had arrived at several methods for filing electronic content on their computers. These had generally developed organically over many years, and organizational practices at one point in a researcher’s career were not necessarily consistent with those at another. Furthermore, the large majority of researchers struggled to find the time to go back and reorganize older files and folders to ensure consistency.¹

Files were almost always organized in the hierarchical ‘tree’ structure of the Windows default system. Popular conventions for determining ‘top-level’ folder structures included: by research topic; by intended research output (e.g. thesis, article, book), often with a chapter-by-chapter approach at the next level down [7] [8]; or by source types (e.g. all journal articles in one folder, notes in another). Non-hierarchical systems were not common, although a couple of researchers

¹ One senior researcher was actually going through her back catalogue of printed and photocopied material, creating PDFs using the faculty photocopier and attaching them to their respective entries in her Endnote library. [6] However, this was unusual.

used bibliographic software to provide an index of their material which permitted searching by keyword [6, 8].

- “For electronic notes, I’ll have a Word document for a particular topic, covering a handful of articles, and these will then be organized into folders. Hard copies of articles are on shelves, grouped by theme.” [3]
- One senior researcher commented that “A good folder structure is important”. She numbered her folders to ensure they stayed in the right sequence, kept separate folders for research and data, and was very careful to ensure different versions of files were clearly labelled. [5]
- A doctoral student who used a lot of archive materials gathered during fieldwork said that she had folders for different types of material: books, archival material, and journals, which were then subdivided into folders each relating to a particular book, archive, or article. [10]

A few researchers described how they cross-referenced between different types of material to help ensure that they had most the relevant material easily findable when it came to writing up research outputs:

- One researcher used the EndNote bibliographic software to associate her notes with particular sources, and even used it as a more general database application – creating her own custom fields (composer, lyricist, rhyme scheme, number of lines, etc.) so that she could conduct searches that would return relevant material. This was an unusually elaborate system, however. [6]
- One researcher with a complex folder system used hyperlinks to connect related information stored in different folders, or to link notes to the PDF of the original document. In general this worked well, though problems were occasionally encountered if files had been moved. [27]

Some researchers wished that they had thought through their organizational structure more at the start of projects so that information was better linked, whereas others found that there were problems connecting sources with notes:

- “Currently I have no way of annotating the images themselves or connecting notes directly to the images. I could convert them all into PDFs, but this would take some time.” [10]
- “Any notes on the content of the PDF are hand-written, and there is no easy way to attach these to the documents themselves – some way of doing this electronically might be useful.” [12]

4.3.2. Organizing emails

A couple of researchers made the point that important research information may be contained in emails, as well as files. This is especially likely to be the case where researchers are working on collaborative projects. When it comes to properly organizing emails, “a lot of people are just stumbling along”, according to one senior researcher [9]. Particular problems identified with emails included topics such as “how much you need to save, how you save it, where you save, and so on” [9].

4.3.3. Organizing paper-based information

Many researchers filed their paper-based materials alphabetically by author, especially printed journal articles. Notes tended to be filed chronologically and divided into separate folders or boxes for different research outputs.

- One doctoral student had tried organizing her paper files (mostly articles) by theme, but this made it hard to find things and necessitated her adding keywords to her bibliographic database, which was extra work. She has since reorganized them alphabetically. This means that she can now find a work by searching for whatever details she recalls in her bibliographic database, which then tells her where to look in her files. [18]
- “I prefer to work with hard copy, as it’s easier to read, so I have a lot of print-outs and photocopies. These are organized in labelled box files by topic.” [9]

It was observed by several interviewees that it’s much harder to find things in hard-copy documents than electronic ones:

- “Paper filing is a long-term headache. Projects that are finished are filed by date of publication, but for ongoing projects it’s harder to know what to do with the material.” [19]
- An early career researcher commented that there is some material that it makes sense to have in hard copy, as reading large sections online isn’t ideal, but it’s harder to keep track of photocopied articles and the like than electronic material, and it’s also harder to find information again. [3]
- “I have several thousand photocopies in my room. These aren’t all catalogued, or even in alphabetical sequence – there are several heaps of papers that have come out of my filing system at some point and never gone back in again.” [28]
- A senior researcher with a lot of hard-copy material commented “in the end, I get what I want, but I just feel very frustrated that it often takes me as long to retrieve it as to write it.” She would like to move towards a system where she has everything catalogued in a searchable electronic index, but this will involve a lot of work. [24]

The sheer physical bulk of hard copy resources was also noted as an issue by a number of interviewees.

- A senior researcher said that she did not have enough file space in her office to store all her papers, so she kept some material at home. This made keeping track of where everything was stored more of a challenge. [5]
- “Keeping piles of articles organized is difficult. Most graduate students don’t have much space, which makes it harder.” [8]
- A graduate student whose research trips involved overseas travel said that she had made a decision to take all notes electronically to minimize the amount of paper she needed to transport. Where possible, she also took digital photographs (rather than photocopies) of primary source material. [10]

4.4. Finding data when it’s needed

The main purpose of organizing research data, whether in electronic or paper-based formats, is to ensure that the most appropriate ideas and examples can be easily found and marshalled when the time comes to write up research outputs. Whilst this may in some cases be a relatively short-term

issue, due to the 'life's work' nature of much humanities research scholars often find themselves needing to return to information recorded many years previously. This adds an additional complication to the challenges of good information organization.

Several of our interviewees reported that they didn't commonly have problems in this regard, rarely losing information or finding themselves in a position where they wasted a lot of time trying to rediscover material that they knew they had. Despite that, however, it was an issue that almost everyone admitted to facing from time to time, and a few quite regularly.

Generally speaking, graduate students were the least likely to struggle to find information that they knew they had when they needed it. However, this seemed to be explained more by the fact that they had simply not amassed as much material over such a long period of time as more senior academics, rather than by any particular awareness of good data management techniques or technological know-how. Those who had only been engaged in academic research for two or three years could remember where they had put things two or three years ago (though it was not uncommon for researchers at all career stages to rely heavily on their memory of where they had put something). The challenge grew with the distance in time and the evolution of personal organization systems.

Comments we received included:

- One graduate student observed that whilst information may get "temporarily lost", "this hasn't been a major problem so far", although she could imagine that it might become one in the future if she pursues an academic career. "I have a large collection of articles and book chapters that I've photocopied or saved electronically, and these are a bit of a mess: they're all kept in one file (organized by author) rather than sorted thematically." [8]
- Another graduate commented that whilst she now had a lot of material it wasn't impossible to keep track of it, but if she continued with her work for a few more years this would ultimately become impractical, and she would have to spend time instituting a better system. "In the future I'd like to have everything in a list or database, as in ten years time I probably won't remember where everything came from." [10]
- "When you've got eight years' worth of data it sometimes takes a bit of hunting." [7]
- "With any filing system, paper or electronic, you need to keep things where it's logical for you to find it – which means there's a risk of problems if someone else had filed something and hasn't put it where you'd put it. It can be surprisingly difficult to find books again when you have a large collection: literary material can be filed under the author name, but this is less useful for more general works, where you're likely to remember the subject matter rather than the author." [9]
- "Organizing notes by project rather than source makes them harder to find." [12]
- "There seems to be an expectation that you will work out an organizational system to manage your data by yourself without any help, and this is hard." [18]
- A senior researcher reported that she had previously arranged notes conceptually (storing all notes relating to one theme in a single file), but had moved to a system where she kept separate notes for individual texts, to improve ease of information retrieval. [20]

The presence of hand-written notes tended to provide a particular challenge:

- “Notes taken on paper can be a nightmare to find again, so I make a practice of typing up any important material so I can search it on the computer.” [17]
- “I usually take notes by hand, but I’m not terribly satisfied with that. It’s fine for something like an article where you’re aiming to write it up within a fairly short time frame, but for something longer, then the chances of it getting lost are quite high.” [29]

Perhaps reassuringly, those senior academics who had spent the most effort developing a systematic method of organizing their data were also those who were least likely to report any problems finding old information. [5, 6]

Those researchers who kept their information in electronic format often made use of the Windows or Mac search functions to retrieve things, negating to some extent the need to maintain a well-organized folder structure. [20]

Many interviewees indicated that if they had known at the beginnings of projects, or even at the beginning of their research careers, what they knew now they would have organized their material differently. [20] However, several researchers also commented that because one cannot always anticipate the direction a research project will take, it’s hard to predict the most useful structure at the outset.

It was observed that there is always a risk with electronic data of the data being lost entirely, rather than simply misplaced. [13] One researcher had lost a paper file containing the basis of a book, but he was relatively blasé about this as he expected it to turn up again eventually. [19]

4.5. Bibliographic software

Many researchers stressed the importance of keeping track of what material they had read, but opinion was divided over the usefulness of bibliographic software to do this. Of twenty-eight researchers interviewed, eight made significant use of some form of bibliographic software, while ten reported that they did not (the remaining ten did not specify). Practice among those who did make use of bibliographic software varied considerably: some used it extensively to store notes, and attached electronic copies of documents to records [6, 8], while others used it just to generate references. More than one interviewee admitted that their collection of references was not up to date [20, 28]. Some of those who did not use bibliographic software had tried it and disliked it [10, 17]; others had thought about using it, but had never got round to it [23]. In some cases, bibliographic software was of limited use for generating references because of idiosyncratic discipline-specific citation conventions (or in one case, lack of established conventions for certain media types) [22, 29].

There was a general consensus that it was best to start using bibliographic software early in one’s academic career, because the later one leaves it, the larger the backlog of references that will need to be entered. Those who did not use formal bibliographic software often had their own way of storing and managing references, most commonly using Word documents, but occasionally in Excel or a database program.

4.6. Structured data

Although there are many humanities researchers that do not work with formally ‘structured’ data, there are a significant number who do. Ten of the twenty-eight researchers who were interviewed

for Sudamih contributed to databases of research information themselves (excluding purely bibliographic databases) and a further eight engaged in some form of data structuring, such as compiling indexes, to help them get to relevant information rapidly in the future. An additional six researchers were intending to embark on a project which would involve the creation of a publicly accessible research database in the near future.

Technological advances mean that it is now feasible to organize and analyse data in ways that would have been impossible in the relatively recent past. This is shaping the sorts of research that academics are engaging in.

- An early career researcher in linguistics observed that much previous work within her particular field (including some of her own publications) has been rather impressionistic in nature. However, the advent of software which can capture and analyse data in a more sophisticated manner has led to a significant shift within the discipline, and one focus of her current work is to use the tools now available to test theoretical claims that have previously been made. [17]
- The project manager for a major online database observed that when people structure their data and see what it's possible to do with it, the direction taken by the research project may well change. [16]
- Two early career researchers who are planning a major database project said that a large part of their aim in doing so was to make it possible to answer research questions which had previously been hard to address. [7]

The creation of structured data plays a wide range of roles within the process of research. Some researchers created small databases purely for their own use: to make the information more easily searchable, or to answer a particular research question for their current work in progress. In other cases, the creation of structured data resources intended for public consumption was in itself a (or even the) major component of the researcher's work.

Unsurprisingly, databases for personal use were generally the work of one individual, although senior researchers were sometimes able to employ research assistants to help with data collection and entry. Research databases constructed for public use often had a team of people working on them (and even researchers working alone on this sort of project had usually sought, or were intending to seek, assistance with the more technical aspects). In some cases, researchers worked on their own database as an individual, but with a view to contributing the finished product to a larger network of related resources, either as part of a formal project team [13], or through informal collaboration [27].

While the collection and organization of data is an essential part of many research projects, it seems generally agreed that data resources contribute less to the academic standing and reputation of the creator than conventional publications. In particular, pure data outputs were not recognized as being of similar value under the old RAE system. The compilation of substantial data resources is time consuming, and consequently may slow the rate at which researchers are able to publish books and articles [14]. Scholars who choose to invest significant time in data projects may therefore find that they are providing a service for others at the expense of their own research career [16].

However, a senior researcher noted that there is a growing expectation that a portion of researchers' time will be devoted to providing just this sort of service: her impression is that within her discipline it is now very unusual for funding to be given for research projects which aim simply to produce monographs, without any kind of spin-off data resource [20]. In some fields (where, for example, historical records or documents are key resources), data-focused projects may aid career progression indirectly, because they result in published editions of primary source material [19].

4.7. Data re-use

The reusability of humanities data tends to depend on whether the data was originally collected with the intention of public use, or whether it was collected to address the specific research questions of the researcher.

When there is a clear intention to make data publicly available from the start of a project (usually via a website), data is usually collected in such a way as to ensure that it is fit for public consumption. Re-use problems tend to arise when data has been collected for personal research, although there can also be difficulties if external circumstances (loss of funding or key personnel, for example) prevent the completion of a data resource which was intended for public use. In general, we found that researchers were happy in principle to share the data they had collected, but there were considerations that restricted the sharing of data in practice:

1. Data may be messy

Unless researchers are thinking from the beginning about making their data publicly available, they tend to compile it in a way that is suitable for their own personal use, but which would not be easy for others to understand or make sense of. The author may be embarrassed to share messy data as it can give the impression that their working practices are not good and the data is not reliable, even though it serves its original purpose adequately. A senior researcher commented that she wouldn't want her personal research database to be out in the public domain without a lot of further work being done on it, as it includes a lot of personal notes which would need to be removed or rewritten, and the time investment this would require was such that she was highly unlikely to do it unless she could secure funding for this project [6]. Another senior researcher, working on a resource which she does ultimately intend to make public, said "I want to be proud of it, rather than being embarrassed by people saying 'You haven't presented this properly'." [24]

Sometimes the available source materials can make it challenging to gather data which is clean and consistent enough to meet standards of public acceptability. In such situations a researcher may be reluctant to commit to publish their compiled data due to inherent and unavoidable messiness.

2. Data employs personal, idiosyncratic standards

When data is gathered for personal use, researchers do not usually concern themselves with adopting recognized standards – they simply adopt a system which makes sense to them in the context of their research. For instance, one senior researcher we spoke to had constructed a searchable database of medieval French songs. Middle French has no standardized orthography, but the researcher in question needed to spell words consistently so that she could conduct searches. She therefore adapted her original sources and changed the spelling of words to render them

searchable. [6] This worked well for her, but would have been confusing, and indeed potentially misleading, to anyone else coming to her database later.

3. Data is partial and specific

People tend just to gather such data as is required to answer the particular research questions they have. Frequently, there is far more information that could have been gathered from the sources which has been ignored and excluded, so as not to 'waste' time. Future researchers are unlikely to come to the data with exactly the same questions as the original compiler, so the partial nature of the data may limit its usefulness. Researchers are aware of how partial their data is and therefore may assume that it is unlikely to be of much interest to others. In some cases, partial data may lead to serious misunderstandings. A senior researcher described a case in which some colleagues had downloaded a dataset from the Essex Data Archive: "It bubbles along at around 96% of children having smallpox [in a particular society], and it then drops to zero in one year, and they said that was the year smallpox was eradicated ... actually, that was the year when they stopped filling the column in – it's got nothing to do with the incidence of smallpox. But you only know that from working in the archive with the real material." [5]

4. The existence of the data is not widely known

Most humanities journals are not interested in publishing the data that underlies the conclusions reached in the articles they publish. If people have based their outcomes on data that they have gathered privately, the existence of the underlying data should be inferable by close readers of the public research, but generally no effort is made to 'market' the existence of the data more widely.

5. The data needs to be milked for publications first

Generally speaking, researchers are unlikely to want to share their data until they feel they have published all the material they wish to get out of it. This is understandable given that – as noted above – traditionally it has been the research outputs in books and journal articles that establish the reputation of an academic and support their career progression. If other scholars were to get access to the data, they might draw similar conclusions from it and then beat the compiler to publication, thus reducing the effective value of the effort in compiling the data in the first place. A senior researcher commented "In principle, you want material to be available, and I believe in sharing. On the other hand, if you've just spent five or ten years collecting a dataset and you haven't yet milked it for what it's worth, and you've had funding to do the project, then you're very nervous about handing over that dataset." [5] Similarly, a graduate student observed that it was vital he maintained control over his data until he has completed his doctorate: if someone else published his findings before that, he wouldn't have enough original material to complete his thesis and get the degree, and several years of effort would have been wasted [13].

Some interviewees had found that possessing a dataset which was *not* publicly accessible brought with it academic advantages. A senior researcher reported that he had been "invited to do half a dozen things" because other scholars knew he had a certain collection of material. He was therefore not eager to make the data public until it could be reshaped into a sophisticated electronic resource, the publication of which would be viewed as having some significant academic merit. Preparing this would be a major undertaking, requiring the collaboration of a team of scholars [28].

6. Political issues may make publication unwise or difficult

One researcher has a private research database which covers a subject area very similar to that of a project with AHRC funding to produce a database for public use. If she were to publish her private database ahead of the AHRC-funded one, this could be interpreted as an aggressive act [6]. Another researcher explained that some scholars in his field (particularly those from Mediterranean countries) are very possessive about material, and that the only way of getting images of or detailed information about certain historical artefacts into an online database would be to allow those individuals to contribute and to be credited as the principal author of that part of the resource [28].

4.8. Storage and backing-up

The majority of interviewees stored the bulk of their electronic data on a personal computer, usually a laptop. Many had more than one computer on which they worked – one at home and one in a faculty or college office, for example, or a desktop machine and a laptop. Some interviewees used synchronization software, but a significant number synchronized their systems manually, usually transferring files using a USB stick or via email.

A minority of interviewees took advantage of faculty or college server space to store their own data [5, 9, 19], but only one of these used server space instead of (rather than in addition to) the hard drive on a personal computer [19]. Large publicly-accessible research databases were generally hosted on servers in Oxford or at other universities, and a few interviewees had previously lodged material with data archives such as the AHDS [4, 12].

Back-up procedures varied widely. The most common method was to use an external hard drive [5, 6, 7a, 7b, 8, 10, 11, 13, 17, 18, 23, 29]. Other methods included USB sticks [3, 9, 20, 22, 26], CDs or DVDs [3, 11], and emailing key files to oneself [9, 23, 29]. Researchers who stored their work on more than one computer sometimes relied on this as a back-up [27]. Some interviewees were meticulous about backing up (one graduate student, for example, had her important folders copied to an external hard drive in Oxford, and to another one kept in London, and had also recently signed up to a commercial online back-up service [10]), but many more were unsystematic and irregular. No one admitted to not backing up at all, though one or two people commented that they probably didn't back up as often as they should.

Less than a quarter of researchers interviewed used the central university HFS back-up service. A couple had tried it but been unable to get it to work properly, or found it was currently impractical to use it as they did not live or work somewhere with access to the university network (this was a particular problem for graduate students and early career researchers). A significant number of interviewees were unaware of the existence of the service. The inability to back up when not connected to the university network (e.g. via VPN) was perceived as a significant disadvantage of the service by a number of interviewees.

4.9. Data ownership and rights

A substantial majority of the researchers interviewed believed that the data they had accumulated in the course of research projects belonged to them, and hence that they were free to take it with them if they moved to another institution.

In fact, according to university policy, any data resources created during University of Oxford employment belong to the university (this is true of most forms of intellectual property, with exceptions for creative works and the copyright on scholarly works such as books, articles, and theses, which remains with the creator). While one or two researchers indicated that they were aware of this [29], it seems that a significant number of researchers are under the impression that data belongs to them when this is not in fact the case.

That stated, a member of Research Services pointed out that in many cases ownership is something of a red herring: the university is highly unlikely to assert any claim to the work unless it has some kind of commercially exploitable value, which is typically not the case for humanities research. In fact, if a researcher who was leaving a university post were to hand over their data before departing, it is unclear what would happen to it: there is at present no real infrastructure to deal with this. In practice, this issue has not frequently arisen [21].

Some scholars who were contributing to major online resources made a distinction between their own personal data and data which belonged to the project [6, 12, 19, 27]. However, the situation was not always clear cut. One graduate student who is part of a major database project team said that while he fully intended to donate the data he is currently collecting to the project, as far as he was aware he was not obliged to do this, and retained the rights to reuse the data elsewhere [13]. In other cases, the rights to the data itself and the database design were owned by different individuals or groups [16, 31]. Additionally, the situation was often complicated by the fact that the content of databases was not always in the public domain. Although image rights belong by default to the person who created the image (or the person or organization who employed them to do so), where databases include images of historical artefacts such as manuscripts or cuneiform tablets, it is not uncommon for some rights over the image to be assigned to the owner of the artefact: hence while the database is permitted to reproduce the images, it may not own them [14, 16]. Another project with a large text encoding element was in the process of navigating a complex rights situation: they were hoping to build on the existing coding of a set of texts which is currently available through a subscription service, raising difficult questions regarding who would own the finished product, and whether they would be able to make the resource available to a wider audience [7]. Some scholars who were in the process of planning the creation of data resources for public consumption expressed some confusion over who would own the material once it was hosted on a university server [3, 7, 9].

A small number of researchers said that whatever the legal situation, they did not feel that they owned the material they had gathered in any morally significant sense [6, 14]. One interviewee said he felt the ideal situation would be one in which no one owned the data, but it was available for anyone to do as they wished with [14]. In many cases, however, desire to share data was tempered by the considerations listed in section 4.6 above.

4.10. Data Policies

The vast majority of researchers interviewed were unaware of any policies relating to the management of their data. Two senior researchers (both of whom sat on relevant committees) indicated that they knew of the existence of university policy documents, but were somewhat hazy about their content [28, 30]. One of these commented that he thought this was probably true of many of his colleagues: "We're aware in a general sense that there are such policy and strategy

documents out there, but quite apart from not necessarily knowing where to look for them, we never get round to reading them ... They end up being very low priority.” [28]

A handful of interviewees mentioned the requirements of funding bodies, generally focusing on what was needed in order to secure funding. Several referred to the need to complete a technical appendix [6, 12, 27] when applying for AHRC funding. One interviewee mentioned the obligation to deposit datasets resulting from UK publicly funded projects with the Essex Data Archive [5], and another was aware of the Oxford Text Archive’s policies (the declarations that have to be made before material can be deposited) [27]. A third had applied for funding from the John Fell Fund, and a condition of receiving this was to check the intellectual property situation with regard to her dataset [29].

A significant number of academics had the impression that the priorities of funding bodies had resulted in a substantial increase in the proportion of data-focused projects within the humanities. However, a member of the Humanities Division research support team suggested that it has always been the case that humanities research projects have usually resulted in datasets of some kind: the chief difference is that researchers are now more likely to view such a dataset as an additional research output and make it publicly available [1].

Those interviewees with a role which involves supporting or advising researchers² were – perhaps unsurprisingly – rather more knowledgeable about policy issues.

Funding body policies mentioned included:

- The AHRC (Arts and Humanities Research Council) is one of the major funders of data projects. Applicants are required to complete a technical appendix demonstrating that they have considered and made provision for the technical aspects of the project. The AHRC also stipulates that the data has to be available for three years after the end of the project, and outputs must be available on open access somewhere. Parts of the AHRC deposit their data with the ADS. [1]
- JISC (Joint Information Systems Committee) also funds a fair number of humanities data projects. They impose requirements comparable to those of the AHRC, but a key difference is that JISC continues to work with projects after the grant is awarded, whereas AHRC involvement tends to be limited to providing the funding. [4]
- The Mellon Foundation is like JISC in this respect. Because the Foundation funded JSTOR, this is their preferred venue for data deposit, although sometimes there are more appropriate UK-based repositories. [1, 4]
- The British Academy has also funded data resources, but doesn’t ask for much technical detail at the application stage (one interviewee commented that they give something of an impression of being technophobes). Once funding has been awarded, they tend to adopt a hands off approach. While they are concerned about the sustainability of resources, they haven’t in practice done much to support or encourage this. [1, 4]
- Leverhulme doesn’t tend to fund data projects: they more usually fund individuals (for example, research leave so an academic can write a book). Like the British Academy, they ask for little information in the application. [1, 4]

² A research co-ordinator, research facilitators, a member of Research Services, and a member of the OeRC.

- ESRC (Economic and Social Research Council) funded projects have to offer their data for deposit, and it has to be presented in a form that the ESRC will accept. They are also thinking of developing an AHRC-style technical appendix. [4]

It was observed that application procedures which force researchers to think through technical issues at an early stage can serve a useful purpose: if there is no incentive for them to do this, problems can result later. (However, even when funding bodies do not ask for this information, the Division's research co-ordinators will encourage researchers to give due consideration to the subject, which helps the project run more smoothly later on.) [1]

More than one interviewee commented that although they may have formal policies relating to data management, funding bodies rarely verify whether these are being followed: while grants are audited from time to time, there is no systematic checking. However, most researchers are conscientious, and their own desire to do the job properly generally provides more motivation to strive for good practice than externally imposed regulations. [1, 21]

Sustainability of data resources is a significant issue. Funders generally don't specify where data should be stored, or how accessible it needs to be: this is very much left to the individual institution. This can leave researchers in a position where they are obliged to preserve their data, but have nowhere suitable to store it. Additionally, as projects usually receive funding for a fixed length of time, there is a serious risk that datasets will be orphaned: there will be a bank of data which needs to be stored, but no funding available for this, and no one with responsibility for looking after it. What actually happens depends very much on the PI and the department in question: some projects have thought this through and have bought server space for future storage, but many haven't. [21]

Provision for long term storage therefore needs to be considered at an early stage of the project. As far as possible, research support workers try to ensure that these costs are factored into the initial project proposal. However, obstacles are presented by some funding bodies, who will not grant funding for any activity which occurs after the official end of the project. [1, 21] In practice, a significant burden tends to fall on university departments, which often lack the financial and technical resources needed to shoulder it. [28] This problem is not limited to those resources created by externally funded projects: as noted above in section 4.8, *all* datasets created during university employment are technically owned by the university, but there is at present insufficient infrastructure in place to support the preservation of these.

Oxford University does not currently have a formal policy on data management. Work is in progress to develop one, though the democratic nature of the institution and the need for extensive consultation means this is likely to be a long process. The policy is likely to suggest retaining data for a minimum of three years, as this would match the lowest of the funders' requirements. In practice, however, data tends to be kept for far longer. [21] The typically long life-spans of humanities data may require particular attention.

A member of Research Services pointed out that it is not sufficient simply to set out the responsibilities of the university and of individual researchers: it is also necessary to consider what the university is going to do to support implementation of the policy. There are some areas in which the university can and does provide support, but others in which the resources to do this aren't currently available, and this would need to be addressed. A policy is of no use to anyone unless it's

backed up with practical guidance. Additionally, it would be important to ensure that the policy is well publicized, and that information about it is available at all levels within the university. [21]

4.11. Conclusions

- Humanities research is hugely diverse, and makes use of an enormous range of types of sources.
- The intellectual value of humanities datasets tends not to depreciate over time – a database of Roman cities is potentially of as much use to researchers in fifty years time as it is today, provided it is not rendered obsolescent through technological change. Humanities scholarship often aggregates to a ‘life’s work’ body of research, with any given researcher often wishing to go back to old datasets in order to find new information.
- There appears to be a growing trend towards structuring data within the humanities. This is partly because technological advances make more sophisticated data projects possible, and perhaps partly because of the changing priorities of funding bodies.
- Most scholars work with a mixture of paper and electronic materials.
- Methods of organizing data also vary considerably, as does the extent to which researchers succeed in creating and maintaining a well-functioning system.
- Good information management is time consuming, and academics often find themselves with insufficient time to keep on top of it.
- However, non-optimal data management also wastes time: a large proportion of researchers report that it sometimes takes longer than they would like to locate material.
- Some researchers find bibliographic software an invaluable tool; others are less convinced of its merits.
- Most researchers are willing in principle to share their data with others, but in practice choose not to do so, for a variety of reasons.
- Most researchers store their electronic information on a personal computer. Only a minority make use of institutional server space.
- Most have some form of back-up routine, although the majority just use an external hard drive and back up on an *ad hoc* basis.
- Central infrastructure services such as the HFS are not widely known or used.
- There is a significant amount of confusion over the ownership of research data. This is exacerbated by complex situations in which multiple people or organizations may have different claims on the same resource.
- Most researchers are unaware of formal policies relating to data management, although those who have applied for funding may know something of the requirements of a specific funding body.
- The University of Oxford currently lacks sufficient infrastructure to support the long-term preservation of research data adequately.

5. Database as a Service (DaaS) Requirements

We asked interviewees whether they could see possible applications for the DaaS in their research, and if so, what features such a service would need to offer to make it attractive.

As detailed in section 4.5 above, a substantial proportion of interviewees were either already using databases or were planning to create databases in the future. In general, the interviewees who showed most interest in the DaaS were those who were about to embark on a database project. With one or two exceptions, researchers involved with existing mature online databases demonstrated little inclination to move from their present systems. There was, however, some interest in the DaaS as a potential means of ensuring the continued availability of Web resources which currently lacked stable, long-term hosting arrangements, and as a tool which would enable datasets which are not presently publicly accessible to be made available online.

5.1. Key database requirements

When asked what a database service would need to offer, a key requirement stressed by many respondents was flexibility. This applied to all stages of database use: database design and creation, day to day operation, and presentation of data.

- One interviewee suggested that the most useful service would be one that could provide researchers with “a bespoke database created from generic tools”. [1]
- “It would need ... the ability to adapt to different sorts of project.” [9]
- “It would need to be flexible – to support long notes fields, and not be restricted to small boxes which can only hold a hundred or so characters.” [10]
- “Flexibility is important – people need to be able to have a variety of fields. Facilities which allow you to export in different formats would be nice.” [27]
- “For me, it would have to be customizable – I’d have to be able to go in and create a layout that did what I wanted. I do that a lot: if [a project worker] has a particular job to do, I create a layout for her that does it, so she doesn’t have to swap between tables.” [16]
- Another researcher said that it would be important to be able to create data entry forms for users. [2]
- “[It needs to have] a usable interface allowing advanced queries and different report formats – including mapping.” [2]
- “Display options – the ability to present your material as you choose – would help to make the service attractive to users.” [4]

In particular, research databases in the humanities need to be able to accommodate a wide range of data types. The ability to link to external sources was also deemed valuable.

- “Humanities research uses a wide range of data types – texts, images, audio, video. For any sort of data you can think of, there’s an example within the humanities.” [1]
- “Being able to link between the database and Web pages (and vice versa) is an important feature of any database service.” [4]
- “Many of our databases contain GIS data in one form or another, and need to be able to interface with mapping software.” [2]
- “It would need to offer the ability to link to external sources, and to include images and multimedia.” [9]
- “It would also need to be able to handle music and art, and hyperlinks to online resources such as digitized collections of illuminated manuscripts.” [24]
- “I’d like to be able to embed or link to images, and to link to entries in a bibliographic database.” [28]

- “It would be useful if it could link to things – so you could include references to other materials that are relevant but not important enough to go into the database themselves.” [29]
- “In time, we’d like to be able link data with research outputs.” [2]

One specific requirement mentioned by a significant number of interviewees was support for diacritics and non-standard character sets.

- “It would need to be usable with a range of languages – with full support for diacritics and other alphabets: Cyrillic and so on.” [9]
- “It would be essential that it was possible to use Arabic text, and not just transliteration, perhaps using Unicode – and you’d need to be able to have right-to-left text.” This interviewee mentioned a useful feature which already exists in Microsoft Access: you can set up a field in which the text automatically appears as Arabic script, which saves time as you don’t have to switch keyboard layouts. [10]
- “I’d need to be able to include different translations... which would include Hebrew and Greek characters.” [24]
- “It would need to be Unicode compatible, or it’s useless to anyone working in a field like Chinese studies.” [27]
- “It would be very useful if it could display Greek texts.” [29]

A fast and sophisticated search function was also a priority.

- “I’d like to be able to create something that was searchable by any variable – so if you have some fragments of information about the person you’re looking for, you can find everyone who matches those criteria.” [5]
- A graduate student working on a major database project specified complex searching and filtering as important functions. [13]
- “Fuzzy searching would be great.” [14]
- “It would need an incredibly fast Web search facility, and the ability to compare information.” [24]
- “Being able to search multiple ways would be useful – so you can search (for example) for things in a particular category, or things within a specific text.” [29]

5.2. Editing rights and security

A number of researchers found the idea of an online database service which would allow multiple people to edit the same dataset attractive. However, this makes it essential that the database service is designed with the preservation of data integrity in mind.

- “A straightforward way of creating relatively sophisticated databases which multiple people can edit would be very helpful – but it’s essential that proper records are kept of what has been edited and who by.” [13]
- Many projects need a database system which multiple people can edit. [19]
- “There needs to be a version control system in place on the server.” [2]
- “Version control is important, because people make mistakes, and you also need the ability to set up editing permissions at different levels.” [14]

- In terms of access controls: *researchers* should be able to enter data, *admins* should be allowed to change the database structure, *anyone* should be able to query the data. [2]
- “Another helpful feature is the ability to restrict input – to say, for example, that a particular field has to contain a year, or a date in a certain format. This is particularly important if you’re working in collaboration, to ensure consistency.” [10]
- “It would need to be secure and properly backed up.” [27]

Careful thought needs to be given to the authentication and authorization system for any such service.

- Researchers working on one major database project commented that using the Oxford Single Sign On (SSO) to control access would not be particularly welcome, as they collaborate with a number of people outside Oxford. [2]

A handful of interviewees saw potential drawbacks to an online database.

- A graduate student commented that he was wary of storing data online, because if the system goes down, you can’t fix it yourself: you’re dependent on someone else, and that person may be away or tied up with another project. [13]
- Security was also a concern: an online database is open to the risk of being hacked into. [13]
- “I’m not sure anything would induce me to stop using SPSS [in favour of the DaaS], because it seems likely that having the database on your own machine and working with it there is going to be quicker than using something based elsewhere.” [5]
- The project manager for a major online database observed that when project staff do not have university office space, access to the database was limited by the speed of one’s broadband connection – which tended to mean accessing it from anywhere other than the building where it was stored was very slow. [16]

Some researchers suggested a combined approach: a system where the data is stored centrally on a server, but can be downloaded for use on a researcher’s own computer.

- “Not all researchers want to work through a web browser, and it should be possible to enable end users to use the database tool of their choice and then upload/download to the server.” [2]
- “If it’s an online system, it would be essential for people to be able to download the material easily and work with it using desktop applications. Allowing collaboration would be very useful, though.” [27]

Where researchers do choose to work online, ease of editing is a key concern.

- “It needs to be easy to edit the database on the server. Lots of people use Open Base on Access at present, so ideally that should be catered for.” [2]
- “It’s important to be able to edit the database quickly: [our current system] struggles with the fact that editing the content requires going to another site and logging in again.” [14]

5.3. Data dissemination

A number of interviewees expressed an interest in the prospect of being able to use the DaaS to make their datasets publicly accessible, rather than simply using it as a collaborative tool for project team members.

- “The option to display material online would be very useful. There are people who are building up private databases, and would like to be able to put them online, but don’t have the funding to do so.” [4]
- “I might be interested in using the service to provide a public interface for my data.” [5]
- “You should never underestimate the abilities of non-specialists in a given area: don’t restrict the data, but instead let people explore it.” [14]
- “A publicly accessible Web interface would certainly be valuable.” [17]

There was also interest in the possibility of using the DaaS to find out what other scholars were working on. The appeal of this feature extended beyond those researchers who were themselves database users.

- “I don’t really work with structured data ... but I would be very interested in a system that allowed me to find out what other people were working on: in my last major research project, it was very helpful to be in touch with other scholars working in related disciplines.” [20]
- “Funding bodies are more and more inclined towards interdisciplinary research, so it’s really important to find out what other people do. It’s easier to do this within faculties ... but if I want to see, for example, who in psychology might be a potential partner for a research project, there’s nothing at the moment that really helps ... I’d be really, really keen on this feature.” [26]
- “Classics is small enough that you tend to know what people are working on, but a central database might provide a way of finding out about side interests you didn’t realize people had.” [29]

However, a minority of researchers expressed doubts about such a service.

- “It does raise a question of what expectations there would be if you put details of your research on such a database: would this be taken as an indication that you were open to answering any queries about your area, and hence are willing to be contacted by anyone and everyone?” [20]
- “Philosophy is a discipline where people tend to work on their own, and they also tend to know what other people are working on ... if I needed to find other people to consult, I’d be more likely to do this by asking my mates – I’m not convinced that having a big database to sift through would help much.” [30]

5.4. Support requirements

More generally, the DaaS as a whole would need to be straightforward to use, and well supported.

- “It needs to be idiot-proof: not something that’s only usable by computer specialists. It shouldn’t be over-complicated, and needs to be intuitive to use.” [9]

- People may need ‘user support’ from the DaaS team when learning to use the new system. [2]
- “Training material for people at different levels and with different levels of comfort with databases would also need to be provided.” [27]
- A number of researchers (especially but not exclusively those who were relatively new to working with databases) also expressed a desire for technical assistance in designing and building systems that met their requirements. [3, 7, 9, 14, 24, 28]

Sustainability was also a concern: it is imperative that any online system should offer stable, secure, long-term hosting and accessibility.

- “Once you’ve put a lot of work into something, you want it to be permanently available, and to have the ability to go back later and add new material.” [9]
- “I’ve had problems finding a stable home for my data: it was on the college website, but got lost in a site revamp.” The data has now been moved, meaning that the URL published in an article which drew on the data is now incorrect. [12]
- A member of Research Services commented that as projects usually receive funding for a fixed length of time, it’s frequently the case that there’s a bank of data which needs to be stored, but no funding available for this. [21]
- A researcher who is also chair of his faculty’s IT committee also observed that there was a long term issue regarding the fate of electronic resources after project funding ends. It usually falls to the faculty to try to pick these up, and while they do have some server space, they don’t really have the in house expertise (or budget) to maintain a large collection of servers. [28]

5.5. Additional features and concerns

Other points raised by researchers include:

- It would be important to have the ability to import existing Access databases. [2]
- Data ownership issues would need to be clarified: if there was a risk that the university would claim ownership of any data hosted on such a service, that would be very off-putting, especially for people who don’t have a permanent job and may want to move elsewhere. [10]
- “The provision of good metadata is important: pretty pictures are nice, but pretty pictures alone are useless for researchers.” [14]

Several researchers expressed a desire for a database service which had features beyond those of a standard relational database. These tended to be scholars working closely with older texts (including manuscripts and inscriptions), who were seeking a way of recording information in a manner faithful to the original, capturing very precise information about textual features, and yet remaining searchable. In some cases the use of XML mark-up schemas such as TEI would be more appropriate for this kind of user than a standard database.

Desirable features included:

- A system that can represent manuscript text – complete with crossings out, interlinings, marginalia, etc. [3]

- A system that can accommodate diplomatic transcript of texts, where variant spellings can be linked to a standard-spelling version of the word. [6]
- A system that can record and present variations between different versions of the same text – different manuscripts or printings, for example.
- A way of tagging linguistic features of audio recordings, so that users can search for specific vocabulary items, case, grammatical functions, tenses, and so on: so that the structure was searchable as well as the actual words. [17]
- A system that can accommodate the text of Greek inscriptions. This would require extensive tagging to make the text easily searchable. Additionally, the text is highly fragmentary, and there may be multiple proposed restorations for the fragments: the system would need to allow users to search as though the words were intact, while still making it plain which words are broken. [28]

There are, however, other projects which have developed or are developing models which will achieve some of these goals. (For example, the KCL-based Inscriptions of Aphrodisias Project <<http://www.insaph.kcl.ac.uk/>> is attempting to apply EpiDoc (a TEI schema for XML markup for inscriptions) to a corpus of material. The MAMA project in Oxford <<http://mama.csad.ox.ac.uk/>> is a second application of this. [28])

A related issue is posed by dates:

- Semi-known chronology data (e.g. Created After; Created Before; Destroyed After; etc.) raises difficulties, and many people don't understand how to structure databases to accommodate this. [2]
- One senior researcher interviewed is working on a project relating to a period when there were multiple dating systems in use in Europe. To avoid confusion, there needs to be some way of relating dates found in texts or other sources to a standard system. [9]

Only one of the researchers we spoke to stated that she integrated her data with other existing datasets (a 'mash up' in the language favoured by technology enthusiasts), [27] although this is a practice that seems likely to increase as more humanities research data is produced and becomes publicly available. Whilst it need not be a high priority of the DaaS at present to cater for mash-ups, it is probably worth designing the DaaS in such a way as not to preclude this in the future.

5.6. Conclusions

- The response to the proposed DaaS was generally positive. While such a service will naturally not be of use to all researchers, a substantial proportion could see some potential application for their own work.
- Interviewees also found other proposed features of the service (such as the ability to make data available online) attractive.
- Researchers with existing, well-established database systems seem unlikely to move to a new system unless they are convinced it would offer significant advantages over their current arrangements. However, a number of researchers in the planning stages of projects were very interested in the service.

- When asked what a database service would need to offer to be attractive, the key concerns mentioned were:
 - Flexibility
 - Web interface
 - Ability to deal with a range of data types, and to link to external sources
 - Support for diacritics and non-standard character sets – ideally Unicode
 - Fast and sophisticated searching
 - Ability to be edited by multiple people
 - Protection of data integrity
 - Ease of online editing
 - Ability to download data and work with it using desktop applications
 - Ability to import existing data in common formats
 - Ease of use
 - Good user support and training
 - Stability and security
- Were the University to assert ownership rights over data stored on the DaaS, this would almost certainly upset researchers and severely reduce take-up of the service.
- Establishing a DaaS requires a very long-term support commitment, as researchers need to be able to cite their data reliably.

6. Training Requirements

We asked each interviewee about their views on ‘data management training’: what they thought was actually meant by this; whether they were aware of training in this area or had even undertaken such training themselves; whether they felt there was any real need for data management training for humanities researchers; and if so what should such training focus on, and how should it be imparted for maximum effect.

6.1. Existing training provision

As well as asking the researchers themselves whether they were aware of any existing research data management training, we also spoke to Humanities Division Research Training Staff, the Divisional Research Support Team, and the University’s Research Support Team, as well as conducting follow-up desk research.

Humanities Division Training

- Introduction to the DPhil (half day course which touches on data management along with a range of other issues).
- Fortnightly events covering key requirements for academic careers, e.g. presentation skills, publishing (journals and monographs), career management, managing the doctorate, and project management.
- Additionally, the Division’s Research Facilitators provide advice for academics putting together funding applications. Where there’s a technical element to the project, this includes facilitating the applicant getting the best possible help with this.

Faculty-Based Training

This varies significantly between faculties, but does not currently seem to provide much coverage of data management.

OUCS IT Learning Programme Courses (<http://www.oucs.ox.ac.uk/itlp/>)

Training on specific software packages includes:

- Microsoft Access (five x three hour courses)
- Microsoft Excel (five x three hour courses)
- NVivo 8 (four x three hour courses)
- MapInfo (three x three hour courses)
- EndNote (three x three hour courses)
- RefWorks (one three hour course for humanities, and one for sciences and social sciences)

The content of the courses ranges from introductory to advanced, and each course is typically offered once or twice each term.

Other courses include:

- Database: Design Essentials (one three hour course)
- ECDL (European Computer Driving Licence) Advanced - Databases
- WISER: Technology tools - reference management (one hour introductory lunchtime session)
- Make: Together: TEI and wikis (one hour introductory lunchtime session)

Skills Portal (<http://www.skillsportal.ox.ac.uk/>)

This offers details of online and face-to-face training available for Oxford University postgraduate students and research staff. The courses are categorized according to the skills listed in the UK Research Councils' Joint Skills Statement (<http://www.vitae.ac.uk/policy-practice/1690/Joint-Skills-Statement.html>), which includes a number of skills relating to data management – in particular, skills C2 and C4 state that researchers should be able to:

- design and execute systems for the acquisition and collation of information through the effective use of appropriate resources and equipment
- use information technology appropriately for database management, recording and presenting information

However, at time of writing, only one course listed on the site covered these skills. This is a course titled Project Management in the Research Context, and while it touches on some issues relevant to data management, this does not appear to be a major focus of the course content.

Skills Toolkit for Research Students (currently under development)

A joint project run by OUCS, the Library Services, and Careers Services, intended to provide research students with an introduction to a range of electronic tools they may find useful.

6.2. What do people understand by the term ‘data management training’?

By this stage of the interviews we had already spent between thirty and fifty minutes talking to researchers about how they managed their own research information and their use of structured data. When we came therefore to asking what people thought ‘data management training’ might encompass, several interviewees defined it according to what they had predominantly been talking about prior to this section of the interview. Interviewees who had been discussing databases at length naturally enough tended to interpret ‘data management training’ as ‘how to design well-structured databases’ or ‘which software tools to use for which tasks’; those who were not users of databases tended to think more in terms of filing and retrieval. A surprising number, however, were quite thrown by this question and needed prompting from the interviewers. It was fairly clear that few of the researchers we spoke to had ever encountered the term before, and that those taking a ‘broad’ approach to data were dealing with a topic that they had not previously considered.

Specific responses included:

- ‘Data management training’ would suggest things connected with data in a narrow sense – the sort of information you might find in a database: tabulated information, or numerical or statistical information. [3]
- Putting things into the Oxford Research Archive might come under the heading of data management. [3]
- Data management training covers “the sorts of things you tell a starting DPhil student: how you organize your bibliography, your notes, the resources that you’re going to need, the bits of photocopying, the conference programmes that you took away – all those bits and pieces of stuff that you might want to look at again. I’d see it as very broad. [It may] ... also involve sound clips, video files, and everything else.” [6]
- Would expect it to be primarily concerned with “number crunching – the sort of thing that happens in disciplines like linguistics where people have thousands of verb forms they want to analyse.” [9]
- It would cover how to manage references. [11]
- It would include storing data, backing it up, and adhering to standards – such as versioning your files. [14]
- “It makes me think of content management systems ... or encoding text according to TEI or other standards, and thinking about documenting your database.” [14]
- Data management tools would include FileMaker, EndNote, RefWorks, and Sharepoint. [15]
- “This would cover everything from just having a big chunk of raw data to how you would then go about organizing it, storing it, backing it up.” [17]
- What sorts of data are worth structuring in a database. “People need to understand the trade-off between the work needed and the usefulness of the result achieved.” [20]
- “I’d take it to mean mainly training in retrieval of data – using databases and online resources to their full potential. It might also cover storing your data in ways that make sense – having a good folder structure and so on.” [23]
- “It would cover classification and categorization of all the materials I’ve got, into a central, reasonable format which means they can be accessed using keyword searches across different functions.” [24]

- “It makes me think of numbers and figures – that’s my first association.” [26]
- “Basically, I think of data as being numbers ... I was an economist once, and that’s what data is in economics.” [30]

A number of interviewees responded to the question with comments on the term itself, instead of or in addition to a definition:

- Any training in this field would need to make it very explicit what it was going to cover and what it would help people to do. [10]
- “I would imagine that many people wouldn’t realize that this sort of training applied to them, as their initial reaction would be that they didn’t use data”. [10]
- If a training course titled ‘Data Management’ were offered, most humanities students would consider it irrelevant to them. ‘Information’ might be a better term. [15]
- “Something I’d like somebody else to do for me!” [24]
- “Data management is quite an off-putting phrase: if I saw a course advertised with this title, I’d probably assume it wasn’t aimed at me, but was intended for administrative staff – it sounds a bit like data entry, only more advanced.” [29]

Given that so many respondents thought of ‘Data Management Training’ in narrow terms – relating specifically to databases – it would be sensible to think instead to referring of ‘Information Management Training’ when offering a broader range of advice. It is also clear that we need to be very explicit about the content of any training provided, to avoid disappointment from those who might understand the offer differently from ourselves.

6.3. Have people already undertaken any data management training?

The exact response given to this question depended upon whether the interviewee was thinking of data management training as being about using databases and software tools or whether they were approaching the subject more broadly. Some had been on courses about using Access or EndNote or other software packages; almost no-one had received any data management training in the broader sense. Asked whether they were aware of the existence of any such training, the answer was a nearly unanimous ‘no’.

- “I’ve read the manual, and I once went to an Endnote training course.” [6]
- Not specifically any training on data management. “It’s something you just pick up as you go along.” [7a]
- “I’ve not had any training, neither on paper filing nor electronic information management” [9]
- “I’m aware that there are courses on using databases, but not aware of anything specifically related to, for example, how one organizes files or other material.” [12]
- “It’s mostly been a case of learning by doing.” [14]
- “I’ve just sort of picked it up, and probably not learnt lots of little tricks that might save time. I’ve picked things up by talking to other people ... but on the whole the people I tend to talk to about research often aren’t particularly technically minded.” [29]
- One research student reported that she had asked more experienced academics about how to organize material, and they had responded with vague mutters about having previously done this using file cards, but they didn’t offer any concrete practical advice. [18]

- This is an issue at other universities as well, and there doesn't seem to be much training available there either. [10]

A handful of interviewees indicated that while they had not personally investigated data management training opportunities, they had the impression that OUCS did or might offer something along these lines.

- "I think there may be a course at OUCS?" [13]
- "I'm sure if I bothered to look, I could find something." [23]
- "I'm aware of the existence of courses at OUCS, and aware that if I needed or wanted I could go on an EndNote training course, for example." [28]
- "I am aware that OUCS run a lot of courses, and I suppose if I were thinking I needed a course on this, I would think to look on the OUCS website and see if there was something." This interviewee (a college lecturer) also said she wouldn't expect this sort of training to be run by her faculty or the Humanities Division, as their courses tend to be focused on research skills for graduate students. [29]

A few interviewees, particularly those with some sort of training role themselves, mentioned existing training that touched on particular aspects of research data management, although all indicated that good data management practice was not the main object of the training.

- For postgraduate students, the Economics and Social History MSc and MPhil courses include a course called 'Tools and Sources', which deals with some aspect of professional knowledge – using online resources, IP, and so forth. It's not specifically a course on data management, but some relevant issues are raised. [5]
- One DPhil student had attended a course on Advanced Training in Linguistics at University College London which included a section on 'getting good data'. This did touch on organizing data, but didn't spend much time on it. [11]
- Another graduate student had received training that touched on data management while a first year undergraduate at another UK university. This covered issues such as taking and organizing notes, and drafting academic work. He commented "It was very useful, because it's not the sort of thing you get taught at school, but university is so much about your relationship with data: if you didn't have it, you'd be a bit at sea." [23]
- The Humanities Division organizes events covering key requirements for academic careers: both things that will help students to complete a doctorate, and things that will be of value to them in their career post-doctorate. They cover things like presentation skills, publishing (journals and monographs), career management, managing the doctorate, project management. There's a course at the beginning of Michaelmas for students in the first or second year of their doctorate which covers a range of things, but includes how they manage their project and their research material. However, this is very brief (a half day course), so it can only touch on these issues rather than going into depth: it's more a matter of alerting people to them than prescribing solutions. [15]

The upshot of this is that whilst data management training in the broader sense is clearly a new field, there are academic skills and methods courses into which it might slot.

6.4. Is there a need for data management training in the humanities?

Most of the researchers we interviewed agreed that there was indeed a need for data management training in the humanities. This feeling tended to be slightly stronger amongst more senior academics who had gathered a great deal of information over the space of many years. Graduate students, who had accumulated less data (and therefore generally spent less time trying to find things), were less enthusiastic although most recognized that there was a role for such training.

- One senior researcher said that she'd be interested herself in such training, as she is thinking about applying for a major research grant at some point over the next couple of years, and this will probably need to include some sort of data-related project. [20]
- As you acquire more and more data, managing it becomes more of a challenge: if you have a rather esoteric filing system, any problems with that are likely to become more obvious as you add more to it. [17]
- "Definitely useful. [...] Many colleagues and others have no idea, no conceptualization of that sort of structure of files, which means they lose things. [...] The desktop will be covered in little icons, because people know they can find it if it's on the desktop. A little bit of training on how files are stored and how you can structure them yourself would be really helpful. [5]
- One graduate student who had been conducting fieldwork abroad commented that she wished she'd thought more about how she was going to keep track of all her information before she went. [10]
- "I can see potential advantages – I'm sure there are useful tricks one could learn, and ways of saving quite a bit of time in terms of not losing references and finding things more quickly." [29]
- "[Data management] is what we do, really, as researchers, and being trained to be a good researcher is always a good thing." [23]
- "I don't know how many people would be keen on doing it [...], but 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 ... so I can see that improving the way I store them might help the actual thinking – apart from saving time, it might be a bit more substantial than that." [26]

There is a possibility, of course, that our interviewees were more enthusiastic about data management training than they would normally be, given that they were being interviewed about the subject by a project that had the creation of training modules as one of its major outputs. There is also the issue that the kind of person likely to consent to be interviewed for Sudamih is more concerned than the average with training provision. Even bearing such concerns in mind, there does seem to be a genuine demand amongst at least some researchers for data management training.

The key problem with training that was recognized by interviewees was one of priorities. With so many pressures on their time, researchers felt that the format of the training and the 'sales pitch' would need to be good.

- A post-doctoral researcher commented that training would be a hard sell. People would probably think it sounded useful, but might not actually do it: partly because it feels as though there are always more important things to do, and partly because "It feels as though

it should be common sense, but I wonder how far common sense gets you – and I wonder how good people are at assessing how much common sense they have.” [17]

- “Most people are so inundated with opportunities to attend training and conferences and workshops that they don’t have time to take up many of them. People tend not to worry about data management until it becomes an issue and there’s something specific they need to do, but even then the usual attitude these days is to try to work it out for yourself on the basis of what you already know.” [6]
- “I’m aware that these things [i.e. training courses] exist, but in the end, for time reasons I end up – probably mistakenly – putting them very low down the list of things that I need to do.” [28]
- “I suspect that probably what would happen is that I might see it and think that would be vaguely useful, but I’m not sure if it’s really useful enough to give up an afternoon ... [but] over the long run, if it saves time, then it almost certainly is worth giving up the afternoon.” [29]
- “I certainly wouldn’t be against it, but I’m not too sure how many takers you would get, just because going to a course takes time, and you can sometimes just read the help files and find out.” [30]

A number of interviewees suggested that uptake might be improved if people did not have to seek out training themselves:

- “The only way you’d get faculties more involved with this sort of thing is to build it into an away-day or similar – something where the training comes to people rather than them having to go to it.” [6]
- A senior researcher suggested that training would be more attractive to graduate students if it happened in their faculty, rather than requiring them to go to OUCS. [27]
- Another senior researcher suggested that training aimed at faculty members would be best channelled through the faculty IT officer or the IT committee, so it was seen as something coming from within the faculty rather than an outside event. “Once it comes from inside, and is potentially supported by the faculty chair, you’re marginally more likely to get some sort of take-up ... It would be the sort of thing not to organize at OUCS, but for somebody to go into the faculty and do it as a session there instead.” [28]
- “If it’s just up on the OUCS website, I won’t get round to looking.” [28]

Several interviewees recommended that data management training be provided to graduate students as part of other, sometimes compulsory, courses.

- One graduate student reported that “a short, fast-moving, mandatory course on general data management skills at the beginning of the course might prove useful”, but that he probably wouldn’t have chosen to attend an optional course. [13]
- “It’s not very easy to make time for this sort of thing. In retrospect, I’d have liked something mandatory at the beginning of the course.” [7a]
- “We’ve just reformed the first year Master’s course to include a section called ‘Sources and Resources’ ... there is no data management component, but maybe there should be, because all graduate students will have some data management needs.” [12]

- “Among some students, there’s an assumption that if they need to know something, someone will tell them: they won’t necessarily go and seek out information” [15]
- “Graduate students are unlikely to choose to pursue this sort of training themselves: people tend to be focused on the skills they need to acquire to progress their current piece of research – not on more general things ... so there might be a case for pushing this sort of training to early graduate students.” [20]
- “I think students would be more receptive to it if they realized how important it was ... Personally, I think there should be stuff built into the course.” [23]

However, those involved in training were not always favourable to compulsory attendance:

- “It babies people, and isn’t consistent with what being a graduate student is supposed to be about – becoming an independent researcher, and figuring out yourself what you need to know.” [15]

A minority of those interviewed were sceptical of the value of data management training. This tended to be on the basis that it would be more useful and effective to provide specific advice as and when needed by researchers rather than generic courses which couldn’t hope to address the diverse issues that researchers have.

- “It’s never occurred to me to send [graduate students] on training courses for this kind of thing ... I’m not a big subscriber to courses.” Instead, this researcher said he provided his graduate students with advice about using specific resources as and when the need arose. [19]
- A JRF suggested that most people wouldn’t be interested in generic training relating to storing articles and so forth, as they have their own methods of organizing material and would dismiss the training as not needed. [3]
- “On the whole, a lot of [graduate students] know a lot more than I do. I don’t think I’ve really noticed any general gaps [in their data management skills].” [30]
- One senior researcher said she thought there was a need for data management training, but that it was hard to identify generic skills that could be taught to graduate students: “I can’t think of one system that would catch them all. Humanities graduates’ theses are as different, one from another, as you can get – you couldn’t make it generic.” [24]

Despite this, many researchers thought that database training was one of the most important aspects of data management training, particularly at an introductory level in the humanities.

- “Many researchers aren’t used to using databases, and instead store their information in Word files” [16]
- One senior researcher specified that he would like training in SQL and advice about good content management systems. [14]

6.5. What should training cover?

Given the divide between researchers who conceived of data/information management in the ‘broad’ sense, and those who thought of it in a more narrow, technical sense, it is probably sensible to maintain this distinction when considering what exactly ‘data management training’ should actually involve.

Although we did not ask the researchers specifically whether they would be interested in learning more about data ownership issues, it became apparent from the level of confusion surrounding ownership and rights that this would be a useful adjunct to any modules we develop.

6.5.1. Information management training

The majority of those interviewees who spoke about broader information management issues, rather than adopting a narrower focus on the technical aspects of designing databases or using particular software tools, agreed that there was room for improvement in this area. However, several interviewees warned that researchers would not be interested in generic data management training. The message was that courses should emphasize the solutions they could provide to specific (albeit widespread) research problems.

- “Training would need to be quite focused – aimed at solving specific problems or achieving certain goals rather than general skills.” [3]
- “[If offered generic data management training], I would probably respond by saying I haven’t got time for that and it’s not for me. But it would depend a lot on how it was pitched ... how it related to people’s problems – for example, are your files disorganized? Can you find things? It’s a matter of relating it to people’s day to day practical issues.” [12]
- “Training needs to be targeted at specific issues, or people simply won’t make time for it.” [14]
- “It’s helpful to be clear about what people are going to get out of a session: data management is a bit of a ‘grey’ term’ [...] How the training is described is crucial.” [15]

Various aspects of information management were raised by interviewees as being important to consider:

- Email management – how much you need to save, how you save it, where you save, and so on. [9]
- “People don’t know how to behave with electronic material [...] I’ve removed myself from all professional group discussion, because there was simply too much information and I don’t have time to process it all. This undoubtedly means I’m now missing things.” [9]
- “In the information age, we do tend to drown in the amount of material ... information management is one way of swimming through all that and organizing it.” [5]
- Organizing paper notes, etc. It might be useful to have a session where people who are experienced in the area could share tried and tested techniques, or perhaps suggest a range of different ways of doing it.” [18]

A number of interviewees commented that perhaps one of the most important functions training could serve was simply to get people to think about how they were going to manage their data before they embarked on a project.

- “It would have been helpful just to have had data management raised as an issue that needed thought at the beginning of the course. Even more helpful would have been some specific case studies or examples of what people have done in the past.” [10]
- “Most graduate students should start their project by thinking about organizing their research data.” [14]

- “People need to think about their data not just in terms of what they want to do now (which is often just to make a list), but in terms of what they might want to do with it in the future.” [16]
- Graduate students need to think about the time that they invest in working with data: if they do this in a systematic way, they can produce something that they or someone else may later find useful. [27]

6.5.2. Technical training on database design and software tools

Several interviewees were sceptical as to whether ‘generic’ information management courses would be worthwhile running, as what researchers needed more was training in particular software or data management methods:

- Training would be better targeted towards learning how to use specific tools rather than acquiring general skills. [3]
- “I wouldn’t seek out training in things like note-taking or generic data management. Dealing with specific technical issues, for example how to run database queries, or how to use relevant software, that would be useful.” [13]

Quite a few interviewees stressed that it would be very useful to be able to get an overview of the various software tools available, with the strengths and weaknesses of each explained, so that people could make an informed choice about what they should use for their own projects. The kinds of software mentioned included database software, bibliographic software, and in some cases data analysis software or collaboration tools as well.

- “One of the key things is simply making sure people know what’s available, and who the right person to go to for help is.” [1]
- “It would be great if there were somewhere historians could get advice about which software package is best suited to their project.” [5]
- “Finding out about ways of having pictures connected to searchable notes would be really useful.” [10]
- “Many researchers aren’t used to using databases, and instead store their information in Word files.” [16]
- “Something people might find useful is a review of different software packages – an overview which covers their advantages and disadvantages and shows what they might be used for ... People think ‘Why would I want to go for two hours learning how to use EndNote, if by the end of that I realize EndNote’s not what I need?’” [15]
- “I would potentially be interested in learning about the software that’s available to help with data management.” [17]
- “It would be useful for [graduate students] to learn to pick the appropriate tool for the appropriate question and the appropriate data ... to know what their options are.” This researcher also commented that it would be helpful for people to know more about open source options. [27]

When it came to database training, we received various suggestions regarding what a course should cover from the researcher’s point of view:

- “It’s worth getting people to think more about recording the relation between data and source. If you’re capturing data created by somebody else, you also need to look carefully at the quality of the data you’re using.” [4]
- Many people approach databases from the perspective of viewing outputs – which does not necessarily help them structure the database sensibly. [2]
- “The biggest danger is that people don’t understand how complex their data is, and so they don’t model it correctly.” [16]

One other aspect of technical data management that was mentioned was the need for researchers to be aware of longer-term curation issues, such as data format migration. The lifetime of a long-term project in the humanities can often exceed that of several generations of software, which means that accessing, editing, and saving older material can sometimes be a problem. In time, data may need to be migrated to different formats in order to ensure continued usability. Factors that can further complicate such issues include the use of proprietary software and standards. [2, 19]

6.5.3. Technical advisory service for Principal Investigators

It was felt that Principal Investigators bidding for project funding might benefit from technical advice. Academics often don’t know where to begin on the technical side of a project, knowing what they want to achieve, but not the steps required to get there:

- “It might be helpful if there were something like a bite-sized one hour course which would give PIs an overview of the sort of questions it was helpful to ask. However, there are dangers in generalizing: this can end up with a course that fits nobody.” [1]
- One senior researcher explained that she was thinking about bidding for project which would have a technical element, but didn’t really know what she would actually need to know about data management to begin, so being able to speak to someone about this would be useful. [20]

The potential problem with courses aimed at more senior academics being too generalized was picked up by others as well, with several interviewees suggesting that an advice service catering for the unique technical problems thrown up by each project would be more appropriate.

- One PI remarked that he would be interested in SQL training and advice on a good content management system, but more personalized help would be beneficial. [14]
- “It would be helpful to teach people data modelling, and how to deal with complex data. They need to understand data structures, and to know how to look at their own data. This ideally needs to be done one-to-one, with someone who knows about setting up databases in that subject areas, who know which questions to ask, and can suggest things that the data owner might not think of.” [16]
- “With many things, if you have any basic knowledge, you very quickly get down to specific problems, so what you really need is potential access to a one-on-one session.” [28]

The IT Learning Programme Team at Oxford University Computing Services reports that many of the people currently attending their database training courses enquire as to whether they could receive more detailed, personalized advice on database design and construction than can be provided in the standard classes. One of the IT trainers estimated that there would be enough call on

such a database consultancy service across all researchers at Oxford to provide work for a 0.5 FTE position.

6.5.4. Training specifically for graduate research students

We asked researchers whether there were particular aspects of data management skills training that were especially important for graduate students, and also what advice they personally would pass on. As might be expected, the more senior researchers generally had clearer ideas about what would be of most benefit, based on their own experiences as their work practices evolved over time. Even graduates and post-docs had usually learnt some lessons regarding good data management during their time, however.

- “Keep track of what you’ve accessed and read via some sort of system like Endnote.” [3]
- “Keep a full history of the data – where it came from and so forth. Versioning is also important [for databases].” [4]
- “Structuring and backing up of files is something students do need to be trained in. The increase in availability of online resources [...] means that people can easily get completely lost in the amount of information, and many students need help in dealing with this.” [5]
- “Back up! You can never have too many copies of your own work.” [5]
- “All students should be introduced to bibliographic software such as EndNote.” [6]
- “Version management is something else that might be useful: graduate students will often have multiple drafts of a thesis, and it’s often useful to have the earlier versions to refer back to.” [7]
- “There are things it would be useful to know at the beginning of a research degree – what you can do with EndNote, for example, and what software is available that might be helpful.” [8]
- “If there’s going to be a user-friendly database service [i.e. the DaaS] ... training in that would be needed.” [9]
- “They probably don’t need to learn SQL, but they do need some technical knowledge: basic training in things like Excel and FileMaker, and XML would be helpful.” [14]
- “There should be a compulsory course where they look at what they’re intending to do in their research, and decide early on how they’re going to manage that data, and how they’re going to present it at the end. People often start off by creating massive Word or Excel files, then two years down the line they have to change everything as they realize that the best way to manage the material is in a database.” [16]
- “Record everything – books you’ve read, lectures you’ve been to, meetings you’ve had with your supervisor and others, things people have said, and so forth. Sometimes something someone has said may click years later, and you suddenly see why it’s important – it’s useful then to be able to go back and check your initial record.” [20]
- “Store as much as possible of your thoughts and what you’ve found.” This researcher suggested perhaps creating a database of your own ideas. [26]
- “[I would advise them] to think about how they can manage their data most effectively and without wasting time taking routes that aren’t ultimately going to be useful.” [27]

6.6. How should training be implemented?

Given that most interviewees felt that there was a role for data management training but that it may be difficult to persuade people to make time for it, the question of how exactly training should be implemented was an important one.

As noted above [see section 6.5.3] there is a feeling that senior academics applying for funding or beginning large data projects would generally benefit more from one-to-one advice than training *per se*. For less senior researchers, or those seeking a more introductory understanding of data management, there was a general agreement that training would be beneficial. There were however two issues that divided opinion: the first concerns whether the desired skills should be taught in the context of division, faculty, or research topic – how much does the subject of research dictate which approaches to data management are the most appropriate? The second concerns the medium of training – is it better to train people face-to-face in a classroom, or offer courses online, or some combination of the two? A third consideration was *when* during the research career is the best time to receive such training, although this did not prompt as much debate.

6.6.1. Institutional context of training

- The training coordinator for the Humanities Division felt that data management training would be discipline-specific, and best handled within the faculties, perhaps as part of students' research methods training. [15]
- A senior Oriental studies researcher commented that within her faculty there were students working on a wide range of materials, and that some of these would have more in common with students in other faculties (art historians, historians, students of literature, and so on) than with many other members of their own faculty – so it might be better to offer training for cross-faculty groups. [27]
- A senior classics researcher suggested that students would be more likely to attend training if it was provided through the existing faculty graduate training programme. [28]
- A senior English researcher suggested distinguishing between subject-specific and more general training. "Different areas [of research] have different needs. There are quite striking differences within faculties [...] but there are also resemblances between groups in different faculties." General data management and research skills, however, might be best integrated into existing structures. [20]
- Another senior researcher noted that the history Master's course had introduced a 'Sources and Resources' component addressing digital literacy, and wondered if data management skills should be introduced as part of this. [12]

It became clear during the interviews that we were unlikely to achieve any form of unanimity regarding where data management skills should be taught. Our feeling, having interviewed researchers across a wide range of subjects, was that whilst there certainly were disciplinary differences, there was also a tendency for these to be exaggerated, and that some aspects of data management could be taught at a divisional level. Care would need to be taken, however, to ensure that examples of good practice were comprehensible and applicable across several disciplines. These assumptions will need to be tested during the pilot training programme.

It is perhaps worth mentioning here that some technical aspects of data management may be so institutionally-neutral and covered well enough elsewhere (on the Web for instance), that it would

not be worth expending resources developing materials specifically for Oxford, but preferable simply to re-use or reference these materials. Likewise, where aspects of data management training are already provided well by other organizations it may be more cost effective to invite those organizations to give sessions at Oxford rather than attempting to recreate their approaches.

6.6.2. Medium of training

Many interviewees observed that both classroom training and online training have their drawbacks. It's not always easy to attend class training with so many other pressures on one's time. The advantage of online training, that it can be done at one's convenience, is also its weakness, in that people can keep putting it off and never actually get around to completing it. Another problem with online training is that one cannot usually ask questions or clarify issues. Pace can be an issue with both forms of training.

- “There is space for both. Face-to-face courses do give you the opportunity to ask questions, so even if the course turns out not to be about precisely what you expected, you may still be able to get something useful out of it. A disadvantage of online training is that people often don't finish it.” [20]
- “I've signed up to online things with all good intentions, and I've only finished them if it's been a requirement for something.” [17]
- “Even if the information is available online, the social and supportive elements of coming along to a course with other people shouldn't be under-rated.” [15]
- “Face-to-face is probably preferable. With online material, it's easy to think it's a great idea but not actually get round to doing it, because there are always other things which are more pressing.” [11]
- “When I've been to courses where there's an expert in the room telling you things, I've found this extremely rewarding.” [23]
- “Online courses can work well as a supplement, but you can't beat face-to-face training: students like it – they like being able to ask questions, and they like training that's tailored to their specific needs ... female students like online delivery far less than males ... People might tell you that they prefer online because it gives them flexibility about when they do it, but the end result is that they won't do it unless it's absolutely vital for the work they're doing. Researchers might do it, but [graduate] students will put it off and never get around to it.” [5]
- “Fitting in another class or course is difficult, given the busyness of an Oxford term – online courses are good!” [14]
- “Online courses can be done at your own convenience, and allow you to pick and choose what's needed.” [12]
- “Online is a good idea, as people can then access things in their own time, and do the training at their own pace.” [9]
- “It's always helpful to have online materials that you can refer back to later.” [8]
- “I quite like the idea of online modules, because you can fit them in when it suits you, and it feels like less of a commitment ... if you did some online modules and felt you wanted more, you might then feel inspired to go to a face-to-face session ... I think I would be much more likely to do the training, or at least experiment with it, if it was online.” [29]

Several of the interviewees who thought about the comparative advantages of different training media came to the conclusion that some sort of combination might actually be most beneficial, using online materials to support a face-to-face course.

- One DPhil student thought that the ideal solution would be “some sort of combination where you could work online, but if you got stuck you could come in to some sort of drop-in office hours and someone would help you”. [10]
- “I wonder if it might be possible to have some sort of brief online introduction which covered the key concepts, but which then led to a face-to-face session.” [22]
- An early career researcher reported that she had previously done some university training (for taking part in the admissions process) which adopted a combined approach. This was a face-to-face course, but there were online exercises that had to be completed beforehand, and feedback was given on them. She felt this worked well: it meant you’d thought through the issues beforehand, and you also got some personal feedback, despite it being quite a large course. [7a]
- Another early career researcher had attended OUCS training, and found this was excellent. The format worked well for him: the face-to-face course involved working through exercises as part of a group, and there was also the opportunity to follow this up afterwards with remote (i.e. online) learning. [7b]

One possible advantage of online training, from a practical point of view, is that it is cheaper than running face-to-face courses.

- Not only are some divisional training course oversubscribed, but they are also currently financed largely via ‘Roberts’ money, a national funding programme that will cease shortly. When this happens, online training may offer a solution. [15]
- A member of Research Services observed “The research that’s been done on this kind of training in the States suggests it’s best delivered in a mixture of formats. Online tutorials are cheap and potentially reach lots of people, but don’t tend to have the same sort of impact as face-to-face training.” [21]

Again, whilst there was no clear consensus regarding the best medium of training, we felt that strongest arguments supported face-to-face training with online supporting materials.

6.6.3. Timing of training

Most interviewees agreed that the best time to confront researchers with the various issues relating to data management was early in their research careers – generally the first two years of doctoral study.

- “Once you’ve built up habits, it’s so much harder to change them. Obviously there are things that you can do to improve them, but if this stuff gets drilled into you at the beginning, you’re going to form your habits in the right way, and that’s just easier all round.” [23]
- “Towards the beginning would be most useful, as you might discover a way of organizing your data that could help you in the rest of your work, and you might regret not having learnt about this earlier.” [22]

Several researchers thought that the second year might be a better time for data management training than the first, as by that point graduate students would be more aware of the issues that they face.

There were some interesting suggestions regarding the timing of training:

- “One of the difficulties with the Oxford system is that there often isn’t time to cover these things in term time, and then at the end of term everyone disappears. Summer schools or courses run in the vacation might be a good idea.” [14]
- “The best time is perhaps when students have been here and working for six weeks or so. By that time they’ve done some essay writing, they’ve had to download journal articles, and have already started to encounter the difficulties of working out what belongs to their thesis [...] Students will learn more one they’re actually in the process of dealing with information.” [5]
- “It might be useful to have the main training at the beginning, but then to have some sort of follow-up, as other issues may arise.” [22]
- “I think that at a later stage, students are more aware, and they become more professional, and they are clearer about their goals ... so I guess they’d be more motivated to make the most of the possible resources available ... so maybe something like the second year of a PhD rather than the first year. I guess it’s also possible to do it as ongoing training, but then I think it would become a bit diluted.” [26]
- “First year graduates are likely to be the most receptive to it, and the ones most likely to do the training.” This researcher also suggested that course for faculty members would be best run either just before or just after the end of full term, “when people are still in Oxford but are a little bit more relaxed”. [28]
- “Doing a doctorate is quite different from doing a Master’s, so you might need some training at the beginning of the first term of the doctorate to cover this ... in the doctorate you tend to be left to your own devices – you can feel thrown in at the deep end and may not be sure if your supervisor is the right person to ask questions ... A one-off event is likely to be more attractive than a course that runs over a period of weeks, as the latter would feel like a much bigger commitment.” [29]

From the interviews we reached the conclusion that it would be most beneficial to catch new graduate students early, but perhaps not right at the beginning of their courses, given that there was often a problem with ‘information overload’ during the first two or three weeks. It may be wise to run courses on the narrower aspects of data management (software choices and database skills) earlier than the broader aspects, as doctoral students and the like are often required to reach decisions about the technical aspects of their work quite rapidly, whereas information management training becomes more meaningful once they have already had the opportunity to assess and start to gather the kinds of sources that their research will demand.

6.7. Conclusions

- Training needs to be based upon actual research problems commonly faced, not promoted as generic skills training
- Training should be provided via face-to-face courses, with supplementary online content

- Graduate students should receive data management training early in their research careers, but ideally not before they have already had the opportunity to assess and start to gather the kinds of sources that their research will demand
- Aspects of data management training should be integrated into existing training where possible
- If possible, information management issues should be included in compulsory training
- If possible, courses should be customizable to allow for the needs of particular faculties. It should be ensured that default examples are broadly applicable across disciplines
- Some attention should be paid to questions surrounding data ownership and rights, as researchers are very vague about such matters at present
- Training should be offered in both 'broad' data (or information) management skills and also in 'narrow' (technical) data management skills
 - 'Broad' data management would include: organizing your files in such a way as you can retrieve information quickly and easily; backing up; versioning; managing email; linking notes to content; keeping track of your sources
 - 'Narrow' data management would include: which type of software is most suited to particular requirements; structuring data in relational databases; querying and retrieving information; long-term curation – data formats, obsolescence, and migration issues; using the DaaS
- An advisory service should be on hand to offer one-to-one technical advice about data management (particularly in the narrow sense) to Principal Investigators wishing to apply for project funding

Appendix A – Interview Question Template

Could you start by telling us a little bit about the research you are engaged in?

- *Do you work in a project team or as an individual?*
- *Do you take notes or sketch ideas in electronic form, or handwritten?*

What sort of ‘data’ do you use in your research? What are your sources?

- *Do you structure this data in any way? E.g. in spreadsheets, tables, by having a defined file structure, specific notebooks for different things? Do you index things?*
 - *If you store data electronically, what software do you use? Are there things you wish you could do but currently can’t, or which are more time consuming/complicated than you’d like?*
- *How is the data then used in your research?*
- *How do you store it? Is it backed up anywhere?*
- *How do you go about finding the right information when you need it?*
- *Do you sometimes go back to things that you haven’t used in a while?*
 - *If so, is it always clear where to look, and what the information means where you find it?*
 - *Do you ever end up ‘losing’ data?*
- *If you were to move to a different institution, what would happen to the ‘data’ you’ve created here?*
- *[if they work on non-Oxford-based data projects] What kind of support do you receive from the institution where the database is hosted?*
- *Would your data be re-usable by anyone else, or is it inaccessible? Is this a conscious decision? Would you object to your data being used by others?*
 - *If there were a central database system which allowed you to find out what other people in the university were working on, is this something you might make use of? Would you be happy making details of your own research available in this way?*
- *Have you ever needed, or would have liked, IT support with your research data?*
 - *What IT support have you received?*
- *Are you aware of any particular policies that relate to your data management, whether these are dictated by Oxford, or by funding councils?*
- *[although not included in the template we were working from, we generally asked people about the proposed ‘Database as a Service’ at around this point in the interviews – most commonly, we would explain roughly what was intended by the service and then asked*

interviewees what features such a service would require before they would consider using it themselves]

Thinking now about training...

- *What would you expect 'data management training' to cover? (Or more generally, what do you understand by the term 'data management'?)*
- *Firstly, have you ever received any training relating to data management yourself?*
 - *Who offered it?*
 - *Was it useful?*
- *Are you aware if any data management training is available? Are there any courses you could take?*
- *Do you actually think that there is any need for data management training in the humanities?*
 - *What aspects would be most useful?*
- *Do you think that there are aspects of data management training that would be particularly beneficial for graduate students?*
 - *When would it be most useful for this to occur? (At the beginning of the course, or later?)*

Appendix B – Index of Interviewees

1	Members of Humanities Division Research Support team
2	Senior Researcher at the Faculty of Classics and members of project team*
3	Early-career researcher at the Faculty of English
4	Senior researcher at the Oxford eResearch Centre, with academic background in medieval and modern languages
5	Senior researcher at the Faculty of History
6	Senior researcher at the Faculty of Music
7	Two early-career researchers at the Faculty of English Language and Literature
8	Doctoral student at the Faculty of Classics
9	Senior researcher at the Faculty of Medieval and Modern Languages
10	Doctoral student at the Faculty of Oriental Studies
11	Doctoral student at the Faculty of Linguistics, Philology, and Phonetics
12	Senior researcher at the Faculty of History
13	Doctoral student at the Faculty of Classics
14	Senior researcher at the Faculty of Oriental Studies
15	Humanities Division Training Officer and Humanities Division website developer
16	Senior researcher at the Faculty of Music
17	Early-career researcher at the Faculty of Linguistics, Philology, and Phonetics
18	Doctoral student at the Faculty of Classics
19	Senior researcher at the Faculty of History
20	Senior researcher at the Faculty of English Language and Literature
21	Research Services Officer
22	Doctoral student at the Faculty of Music
23	Doctoral student at the Faculty of Theology
24	Senior researcher at the Faculty of Theology
25	Early-career researcher at the Rothermere American Institute
26	Early-career researcher at the Faculty of Philosophy
27	Senior researcher at the Faculty of Oriental Studies
28	Senior researcher at the Faculty of Classics and Chair of IT committee
29	Early-career researcher at the Faculty of Classics
30	Senior researcher at the Faculty of Philosophy

31	Senior researcher at the Faculty of Oriental Studies
----	--

* Two separate interviews were conducted with the researchers on this project, one involving the senior researcher and other members of the project team, the other with those members of the project team with responsibility for database maintenance.