Researchers' Database Behaviour and Use: A Literature Review

Virtual Infrastructure with Database as a Service (VIDaaS) Project

vidaas.oucs.ox.ac.uk

Authors Luis Martínez-Uribe
Meriel Patrick

Affiliation Oxford University Computing Services

JISC

August 2011

Researchers' Database Behaviour and Use: A Literature Review

Contents

1. Introduction

This literature review was compiled in the summer of 2011 as part of the VIDaaS Project, funded by JISC and HEFCE under the University Modernisation Fund. Its chief aim is to survey the literature on the academic use of databases and allied technologies (including one of particular relevance to the VIDaaS Project: cloud computing). It covers both journal articles and the relevant grey literature – specifically a number of reports published by interested bodies.

The literature on general database use in academia is surprisingly sparse, and much of what does exist is focused on the sciences. To round out the picture, therefore, Section 3 includes a number of case studies of individual database projects, including three with a humanities focus. Section 4 offers a very brief overview of some of the guidance available for researchers.

2. General works about technology use

2.1 Database technology

2.1.1 In the academic sphere

The issues that surround the collection, accessing, and sharing of data are perceived as significant challenges in many research domains. Better use of technology and the implementation of information management strategies have the potential to improve many processes, and thereby to help move research fields forward.

Databases in particular can form a key part of an information management strategy which can support researchers and facilitate data management throughout the research lifecycle.

- Database technology can assist at the data collection stage by providing customized ways to input data, and the ability to do so from different locations (Burns, Fincham, and Taylor 2004, especially pp. 731-4).
- If harmonized protocols are used at the design phase, databases can enable integration with other datasets in the same field (Cushing et al. 2007, *passim*).
- Databases can help with the management of an ever increasing corpus of heterogeneous digital materials including documents, pictures, spreadsheets, etc. (Evrenosoglu 2008, p. 2).
- Analysis of the data can also benefit, through automatic database calculations and reporting (Evrenosoglu 2008, *passim*), plus data visualization (Cushing et al. 2007, pp. 18-19).
- Access to data can be improved by better and more flexible ways of searching (Buurma, Levine, and Li 2011) and publishing the data (Johnson 2005, p. 100).

However, databases can be complex, and the technical skills required to develop and maintain them represent a barrier to researchers (Cushing et al. 2007, pp. 9-10). Some strategies to overcome this barrier include empowering researchers to do the database design with tools that simplify the process (Cushing et al. 2002, pp. 11-17; Cushing et al. 2007, especially pp. 11-18) and involving multidisciplinary teams where IT specialists, librarians and researchers are brought together (Buurma, Levine, and Li 2011).

Data sharing and curation are also prominent topics in the literature examined. Since the OECD seminal report “Promoting access to public research data for scientific, economic, and social development” (2003), there has been energetic discussion on sharing of and access to publicly funded research data, and the effects of this on research and innovation. Data from publicly funded research is viewed as a public good, and as data is central to the research process, it is desirable that it should be made as widely available as possible. Sharing issues are therefore international in scope. Researchers, institutions, and funding agencies worldwide have a role to play in cooperating to address the need of sharing and access to research data (p. 4).

The OECD report observes that access to research data reinforces scientific enquiry and promotes new research (p. 2). Goth (2010) makes a related point: there are areas where the scope and complexity of the data makes it virtually impossible for any given research project to mine its full potential. Making the data widely available

therefore opens up opportunities for more researchers to work on it, and for more results to be obtained from the same dataset (pp. 14-5).

Nevertheless, there are also some barriers to sharing research data. Via a survey of ecology researchers, Cushing et al. (2007) identify several, all of which seem likely to apply in many other research domains. In practice, researchers often want to exhaust their data's usefulness before releasing it, and may also have concerns about misinterpretation of the data. Hine (2006) notes that retaining control over data is a particularly pressing concern for researchers at the beginning of their careers, who need data of their own to complete a doctoral thesis, or to gain recognition via publications (p. 284). In the database project she studied, this issue proved to be something that had to be taken into account by the technical staff: "An important aspect of the database design was to incorporate appropriate safeguards to ensure that those contributing data would feel that they were being adequately protected against misuse of their data and were receiving credit for their work" (p. 286). The provision of differing permission levels for different users was one important means of achieving this (p. 287). She also describes an interesting exchange mechanism, whereby users of the database were only provided with the data they needed when they contributed information of their own, thereby encouraging further submissions (p. 288).

An additional barrier reported by Cushing et al. is that preparing data for publication is time consuming, and there are currently no obvious personal rewards for doing so (2007, p. 9). This combination of a lack of time and of motivation is also picked up on by other authors. Westra (2010) observes that while some of the researchers who took part in a needs assessment for scientific data services showed an interest in publishing or depositing their data, in practice the task of dealing with data generated by current research usually took precedence. A DCC-commissioned report (Key Perspectives, 2010) notes that even in projects which had a dedicated data management support service in place, adherence to the data management policy (designed to generate datasets suitable for preservation and reuse) was patchy: 'The traditional view that the primary outputs of a project are journal articles prevails, with the fate of the underlying data coming further down the list of priorities' (p. 9).

On a more positive note, while many of the researchers' data management priorities identified in Westra's needs assessment (2010) have current research as their immediate focus, the list includes a number of things that are likely to aid data curation and sharing in the longer term. Among the researchers interviewed, the top priority (by a substantial margin) was data storage and backup. Others included making data findable by others; allowing and controlling access to the data; connecting data acquisition to data storage; and documenting and tracking updates[1]. Comments elaborating on researchers' selections from the priorities list expressed a desire for help keeping data organized, coherent, well documented, and in a form that would facilitate future retrieval. It seems that database technology which meets researchers' needs for their day-to-day work may also have benefits later in the research data lifecycle.

And despite some apparent ambivalence to data curation, when the work needed to preserve data has been done, most researchers seem keen for it to remain available for

[1] Westra's article provides useful summaries of the researchers' priorities in graph and tabular form.

the long term. The Sudamih Researcher Requirements Report (Wilson and Patrick 2010) observed that humanities datasets tend to retain their intellectual value over time: they do not typically go out of date, and researchers often wish to revisit older datasets to find new information (p. 10). Westra's needs assessment (2010) reveals that in this respect scientists seem to have more in common with their humanist cousins than one might perhaps anticipate: for 30 out of 34 data assets surveyed, their creators believed that the data should be preserved indefinitely.

The Key Perspectives report (2010) notes some specific challenges that must be addressed in the realm of curation. Disciplinary differences in data type have resulted in the development of a wide range of technical standards and formats. This makes the task of curation in multi-disciplinary repositories (such as those attached to a particular higher education institution) a difficult one (p. 3). Furthermore, in fast moving fields (such as those involving video data), preservation requirements may evolve within the course of the project, requiring a level of technical support that few institutions are able to provide at present (p. 10). Specialist data centres may provide an answer, but researchers do not always have access to these. (Although the focus of the Key Perspectives report is on data curation and reuse, one might also note that the variety of data types and rapid evolution of projects pose corresponding challenges for the database systems or other technology used to store and analyse material throughout a project's life.)

Once data has been shared, there may still be bars to reuse if it is not available in a form that allows researchers to make use of it easily. The Research Information Network Report *Reinventing Research?* (2011) notes that researchers expressed a desire for 'more uniform standards for digital archives and cross-referencing capabilities across datasets' (p. 61). This sentiment is echoed elsewhere in the report: connecting digital archives to each other has the potential to simplify the research process dramatically (p. 29; p. 35). Busy academics have a finite amount of time to spend on searching for material, and hence are unlikely to be able to consult dozens (or perhaps hundreds) of discrete sources: the chances of locating relevant material are vastly improved if databases can be linked, or made cross-searchable.

An article by Christine Hine (2006) considers a different aspect of the topic: she conducted an extensive ethnographic survey of a particular scientific database project (one focused on mapping the mouse genome), with a view to examining the impact of the rise in the use of databases on contemporary scientific practice. While she dismisses claims that an increased focus on databases will ultimately relegate lab experiments to the history books and transform biomedicine into a pure information science, she acknowledges that databases do bring changes to the frameworks used for evaluating research, researchers' work practices, and (partly because of the increased potential for multi-site collaboration) the spatial environment in which the work occurs (pp. 270-1). She cites other studies which have demonstrated that information technology can have a significant – and often unpredictable – effect on the way research is done, and notes that the work that is done in this area today may influence the sort of work it is possible to do later: "In creating a large-scale infrastructure, database designers may be shaping the future possibilities, not just of their databases, but also of the science that revolves around them" (p. 273).

However, this does not imply that the use of databases will inevitably have an impact on the way that the research material itself is conceptualized: in the project she

studied closely, while co-operating with the database developer pushed the scientists to provide a more explicit model of the information (that is, the rules it conformed to, and the possible exceptions to those rules) than they might otherwise have worked with, both developer and scientists agreed that the purpose of the database was simply to represent (rather than create or reshape) an existing natural ordering (p. 278). Hine ultimately concludes that "The digital perspective... may shape goals to a certain extent, but it does not determine outcomes... while practices and outcomes of knowledge production may change with increasing use of information and communication technologies, such changes do not do away with existing frameworks" (p. 292).

Hine also makes an interesting observation about the way that scientists and database developers work together. For the project to succeed, a significant degree of mutual understanding (both technical and cultural) was necessary. However, it was far more important for the database designer to understand the nature of the scientists' work than vice versa: building a useful database system required a good grasp of what the scientific work involved, whereas in contrast, the scientists needed only to know how to operate the finished system, and not the details of how it was constructed. This disparity did lead to occasional difficulties, however, particularly in terms of the perception of the scale of problems: the scientists had little conception of whether a requested modification to the system was trivial, or one that would require some weeks' work. This led to some frustration on the part of the database designer (pp. 281-2).

2.1.2 In the wider world

Database use is not, of course, limited to researchers. A full survey of the more general literature on database technology falls outside the scope of this review: however, this section draws together a few key insights that may also be of particular relevance within the world of academia.

Margo Selzer (2008) observes that complexity is increasingly becoming an issue in the world of relational databases. Relational database management systems (RDBMSs) now offer an enormous (and still growing) range of features. This has an impact both in terms of price and personnel: it is often necessary to employ a database specialist to manage the system (p. 53). However, the likelihood is that any given use of a system will actually employ only a small fraction of the available feature-set. 'If an application doesn't need functionality,' Selzer argues, 'it should not have to "pay" for that functionality in size (footprint, memory consumption, disk utilization, and so on), complexity, or cost' (p. 55). The solution, she suggests, is a combination of modularity and configurability: it should be possible for users to select only the components that they need, and then to configure the system to meet the needs of their particular operating environment (pp. 55-6).

The Claremont Report on Database Research (Agrawal et al. 2009) makes a related point: while traditional database management systems are still popular, there is an increasing trend toward the development of new data management solutions, custom-built from simpler components (p. 57).

An interview with Pat Selinger, pioneer of relational database management systems, follows Selzer in noting cost as a significant issue (Hamilton 2008). Selinger, however, observes that it is not principally the cost of the technology itself which may

be the major consideration in years to come: the cost of processors and disc space is in fact decreasing. The cost of labour, however, is going up. When combined with the probable rise in the sheer quantity of data that needs to be taken care of, this results in an urgent need to find new ways of improving efficiency in data administration and curation, and of enabling each individual working in the field to handle vastly more data than is currently possible (p. 34). This point echoes sentiments expressed in the academic literature: for example, the DCC-commissioned 'Data dimensions' report (Key Perspectives Ltd 2010) observes that in the life sciences in particular (where a single experiment may use thousands of data files, each with millions of rows), the quantity of data produced is growing at a rate which is 'largely outstripping attempts to curate it' (pp. 10-11).

Size is not the only feature of datasets that is currently in flux. The Claremont Report draws attention to the fact that the *type* of data that is being used is changing: unstructured and semi-structured data are becoming increasingly prominent features of the information landscape. The increase in the quantity of data accessible online also means the information to be managed may be far more widely distributed than was the case in previous decades (Agrawal et al. 2009, p. 61). It is clear that the data management challenges of the future – both in academia and outside it – extend far beyond those of dealing with the sort of information that can be stored in a traditional RDBMS.

2.2 Shared services and the cloud

While databases are an immensely powerful tool with much potential to enhance academic research, they are nevertheless only one piece of the jigsaw. In addition to advances in the software used by individual researchers or research groups, recent years have also seen significant developments in the type of IT infrastructure that HEIs are aspiring to – or are already providing. A notable aspect of these developments is the rise of shared services and cloud computing.

Shared services are those provided to a related group of customers by a service team connected to, yet remaining distinct from, the customers themselves. In a higher education context, this might involve selected IT services being available across a group of HEIs, rather than each institution having to make their own arrangements. The cloud computing model allows customers to pay to use, rather than to own, computational resources – so, for example, an HEI or research group may pay to use a centrally-managed data storage system, rather than maintaining their own servers. That these strategies are viewed as having the potential to be key components of the information landscape of the future is evidenced by HEFCE's and JISC's major investment in the UMF Shared Services and the Cloud Programme[2].

The Clark, Ferrell, and Hopkins report (2011) emphasizes the importance of shared services in the current UK economic climate: they have an important role to play in making universities more competitive in the future, whilst also gaining significant savings (p. 1; p. 27).

Two features of cloud computing are enabling rapid progress and uptake: first, it is based on existing mature technologies which are already widely used, and secondly,

[2] See http://www.jisc.ac.uk/whatwedo/programmes/umf.aspx for more details.

‘the cloud’ is conceived of in sufficiently abstract terms to enable interested parties to buy in at a range of different levels (p. 14). Two obvious targeted services are data storage and hosted email (p. 17).

Some of the potential benefits include improving the use of the IT real estate across an organization, providing resilient, cost effective architectures, scalability, enhancing agility, and improving economies of scale (p. 21ff).

The report also provides a useful categorization of shared services as cloud models (Appendix 1) that includes the following:

- Infrastructure as a Service (IaaS) – includes processing, storage, networks and other computing resources
- Platform as a Service (PaaS) – applications within a computing environment
- Software as a Service (SaaS) – software environment maintained by the provider

There are some more cautious voices, however. Willcocks, Venters, and Whitley (2011) warn against the dangers of buying too heavily into the hype. A move to cloud-based computing is likely to have significant consequences, and many of these are not yet completely understood. Legal compliance and the management of contractual relationships between client and cloud provider are among the challenges that the cloud is facing (p. 3-4).

The Claremont Report (Agrawal et al. 2009) anticipates that cloud-based services will become increasingly prevalent, and notes the benefits that cloud computing offers in terms of cost saving (p. 62). It also, however, details a number of challenges that need to be addressed. For example, the cloud data services available at the time of writing typically offered limited functionality when compared to traditional database systems, and existing systems did not always scale effectively when employed in the context of an extensive shared infrastructure. Additionally, the limited level of human intervention in a typical cloud system means that a greater proportion of the work must be automated (though if this can be done successfully, this is of course one of the features that enables cloud-based services to offer efficiency savings), and when data is no longer confined within physical boundaries, security becomes a much more pressing consideration. On a more positive note, the report observes that awareness of these concerns is in fact helping to accelerate development in these areas (pp. 62-3).

3. Specific database case studies

The following sections offer a number of case studies of particular research projects where database technologies play a key role. Particular areas of interest described here include the data management problem that drives the development of databases, barriers to data sharing, and particular database features and functionalities developed to solve initial problems and support research processes.

3.1 Databases in delay analysis

Complex projects of all types that make use of the critical path method require a comprehensive delay analysis in order to have a handle on the progress made against the plan. Evrenosoglu (2008) argues that in spite of advances in scheduling software, little has been done to support the processes involved in delay analysis. A major issue

is that the processes combine a variety of documents and software to support the work of the analyst. Most of these tools are not integrated with each other, which results in the manual input and updating of the information over and over (p. 2).

The author proposes a relational database solution that would serve as a repository for all the data and information contributing to the analysis, with useful automatic calculations and reports. The data entry can be then delegated to administrative staff, whilst the analyst focuses on examining the data (p. 1). The pilot database developed imports all the scheduled activities, allows filtering of those with finish dates later than the planned date, produces reports with the activities and charts displaying the variances, and allows logging critical documents to improve search and access (pp. 2ff.). In sum, the database solution does not supplant the expert analyst, but by managing all the information, makes the analysis easier and more efficient.

3.2 Canopy Database Project

Cushing et al. present in two papers (2002; 2007) the Canopy Database Project for researchers in ecology (particularly those engaged in forest canopy research). As in many other fields, data collection occurs individually, and generally researchers maintain their databases in spreadsheets and flat files in their own computers (2002, pp. 3-4; 2007, p. 8).

The project undertook an initial survey of researchers in ecology, and discovered that collecting, using, and finding data are perceived as the biggest challenges they face (2007, p. 11). Several major barriers to data sharing are also identified. One is that documenting the data is time consuming and there are no rewards; others include the issues of wanting to glean all usefulness before sharing with others, misunderstandings about citing data, or fear of the risks of misinterpretation of data (2007, p. 9).

The authors identify information management, and database technology specifically, as one of the practices that will help move forward the field of ecology research. An interesting point of the paper is the recognition of the importance of empowering users to do some of the programming with appropriate tools. The Canopy Database Project developed a set of tools to support the management of data by canopy researchers. One of the tools aimed to help researchers to design their own databases, by simplifying the process with the provision of common structures (templates) that use domain-specific database components. This should empower researchers to develop their databases within a framework that will potentially allow the ultimate integration of datasets. Other features of the database developed included easily customizable data entry forms and visualization functionalities (2007, pp. 11-19).

3.3 Clergy Church of England Database

The Clergy Church of England Database (CCED) project was a five year initiative funded in 1999 to create a relational database of clerical careers in the Church of England between 1540 and 1835 (Burns, Fincham, and Taylor 2004). The process involved overcoming challenges such as the geographic dispersion of the paper accounts among 28 archives, the variety of the ecclesiastical records, and the linkage of records (by person, place or bishop) (*passim*, especially pp. 727-30).

The final database allowed inputting data from dispersed locations, modification of the initial schema as the project progressed and the organization of the data was better

understood, and linking related records to track individuals' career moves across diocesan boundaries (pp. 729ff.). The authors mention the need to justify its purpose and usefulness to sceptical colleagues, so they include examples on how the database could bring important new insights of the clerical communities (pp. 726-30).

3.4 Early Novels Database

The Early Novels Database (END) is a bibliographic database of 3000 novels. Buurma, Levine, and Li (2011) explain how with the proliferation of massive digitization projects such as GoogleBooks, important information about specific editions and copies has been lost. This affects scholars who want to locate particular editions of novels.

The END was produced through a collaboration between librarians, IT specialists, faculty, and undergraduate researchers at the University of Pennsylvania. The project also used graduate students to create records in the database. The students received training, and many of them had a personal project related to the database work. The database project also provided an opportunity for the students to work with IT staff, librarians and professors.

The functionality of the database highlighted as most useful is the faceted search. This facilitates the work of researchers as it helps them to locate and view the information in more flexible ways. One of the potential values perceived by the authors is the possibility of inspiring other forms of collaborative humanities research.

3.5 Old Bailey Online

Old Bailey Online is an extensive Web resource, offering records of proceedings from London's central criminal court between 1674 and 1913. The Research Information Network Report *Reinventing Research?* (2011) offers some reflections on researchers' use of the database.

A typical pattern involves the use of keyword searches to locate relevant material, then saving portions for personal use: most of the users interviewed maintained 'some sort of mini-database' on their own computers. Programs used to manage the information included Microsoft Word, Excel, and Zotero (p. 24). Most respondents needed to annotate or make notes on the text in the course of their work. A variety of strategies were employed, including copying passages into Word and using the comment and highlight functions (p. 25).

While the report does not indicate that the scholars using Old Bailey Online were dissatisfied with it, it does say that many of them felt 'that their organisational strategies are haphazard and project-based' (p. 24). The usage patterns reported highlight the fact that when researchers are making use of material drawn from a publicly accessible source, they often wish to reorganize and add to that material. At present, Old Bailey Online offers few tools for doing this, and the result is improvised and 'haphazard' research practices. This reflects the widespread need for customizability: if public data sources are to provide maximum benefit to researchers, they may need to be designed with this in mind.

4. Advice and guidance for researchers working with data

Given the barriers to using databases or sharing data identified throughout this literature review, it is clear that researchers stand to benefit from advice and guidance for working with data. The UK Data Archive has devoted much attention to the management and preservation of social science data for over 40 years, and the second edition of their data management guide (Van Den Eynden et al. 2011) provides extensive best practice advice. The guide follows a data lifecycle approach, with sections covering data management planning, documentation, format and storage, ethics and consent, plus copyright issues.

The reasons for sharing data – such as transparency and accountability and the effect on innovation – have driven funding agencies and journals to require data sharing (p. 3), and this puts not only the researchers, but also their institutions, on the spot. Consequently, research institutions need to be able to support their researchers in their data-centric work, to enable them to fulfil their obligations to funding agencies. The OECD report (2003) proposes a framework for analysing the data challenge, with five domains that can help research institutions to assess their readiness (pp. 8-11):

1. Appropriately designed technological infrastructure
2. Variety of institutional models that support the needs of researchers
3. Continued and dedicated financial support
4. Legal domain to support access and sharing
5. Reward structures to promote access and sharing

5. Summary and conclusions

It is clear that database technology offers both many opportunities and many challenges for the academic community. The increased sharing of data that databases (particularly online systems) can facilitate also has the potential to have a substantial positive impact on research practice, as do up-and-coming technologies such as cloud computing. Some of the benefits and some of the difficulties still to be overcome that are highlighted in the literature surveyed are listed below. These lists are not, of course, intended to be exhaustive.

5.1 Benefits and opportunities

5.1.1. Benefits of database technology

- Customized methods of inputting data for different contexts allow diverse bodies of information to be captured
- The ability to input data from a variety of locations
- The possibility of integration of multiple datasets
- Easier management of large collections of heterogeneous materials
- Improved or easier analysis of data, via automatic calculations, reporting functions, etc.
- Tools which permit data visualization
- Improved and more flexible ways of searching data
- Improved possibilities for collaboration, both between different locations and across disciplinary boundaries

- Access to and rights to edit data can be controlled via differing permission levels
- Changes to data can be documented and tracked
- The construction of a database may help make explicit the conceptual model to which the data conforms
- In some cases, data entry can be delegated, freeing up researchers' time to concentrate on the analysis
- Duplication of effort may be reduced
- With appropriate tools and templates, researchers can be empowered to do at least some of their own programming

5.1.2 Benefits of data sharing

- Improved transparency and accountability in the research process
- Improved access to datasets
- New research is promoted
- A better return on the investment of public money, as shared data may be reused multiple times
- The possibility of more efficient use of large datasets that are too vast or complex for a single research project to mine fully

5.1.3 Benefits of shared services and cloud computing

- Significant cost savings through economies of scale
- Universities are aided in becoming more competitive
- Improved use of IT real estate across an organization
- The provision of resilient, cost effective architectures
- Services can be scaled to users' needs
- Enhanced agility and flexibility
- Automated procedures may enable researchers to do more without assistance from IT support staff
- The specific challenges of cloud computing are helping to accelerate development in some areas (e.g. security)

5.2 Challenges remaining

5.2.1 Challenges of database technology use

- Developing databases requires a set of technical skills which researchers may not have or wish to acquire
- The complexity of modern RDMBSs means they are often expensive – in terms of the initial purchase, the hardware needed to run them, and the personnel needed to maintain them
- Many projects require (or would benefit from) a degree of customizability that is currently not easily available
- Long-term curation of datasets
- Effective and efficient curation of vast (and growing) quantities of data
- Management, analysis, and curation of data conforming to a wide range of different technical standards
- Management, analysis, and curation of unstructured and semi-structured datasets

- The full effects of introducing databases or other technology may be unpredictable
- Researchers and database developers may lack necessary knowledge of each other's working practices

5.2.2 Barriers to data sharing

- Even when facilities for sharing data are available, researchers may be wary of using them
    - Because they want to exhaust their data's usefulness before releasing it
    - Because of concerns over data misuse or misinterpretation
    - Because of concerns about not getting full credit for their work
- Preparing data for publication is time consuming, and there are currently few personal rewards for doing so
- Dealing with more pressing day-to-day tasks often takes precedence over planning for the long-term fate of a dataset
- If shared datasets are not linked or cross-searchable, there may simply be too many sources for researchers to access
- If shared data is not provided in a suitable form, or with suitable tools, it may be less useful to researchers

5.2.3 Challenges for shared services and cloud computing

- The possible consequences of a move to cloud computing are not yet fully understood
- Legal compliance
- Managing relationships between users and cloud providers
- Current functionality may be more limited than in other computing models
- Existing systems may not always scale effectively
- A larger proportion of operations must be automated
- Data security

5.3 The way forward?

This final section draws together some themes from the literature surveyed, and suggests some areas of focus which may enable academic IT service providers to best meet the needs of their users.

- Tools which empower researchers to build and program their own systems
- Enhanced understanding of users' requirements and work practices
- Database systems that can accommodate diverse material types
- Harmonized protocols to facilitate integration of datasets
- Safeguards to protect shared data against misuse, and to ensure the creators are properly credited
- Tools to assist researchers in keeping data organized and properly documented
- Tools to make data findable, and to aid efficient data retrieval
- Systems which aid, encourage, and reward data sharing
- Modular systems which allow users to select (and pay for) only the functionality they need
- Flexible and easily configurable systems

- Methods of allowing fewer people to administer and curate a larger quantity of data
- Sufficient advice, guidance, and support for researchers

6. Bibliography

Agrawal, Rakesh, Hector Garcia-Molina, Johannes Gehrke, et al. 2009. The Claremont report on database research. *Communications of the ACM*: 52(6): 56-65. Available at: http://portal.acm.org/citation.cfm?doid=1516046.1516062.

Burns, Arthur, Kenneth Fincham, and Stephen Taylor. 2004. Reconstructing clerical careers: The experience of the clergy of the Church of England database. *The Journal of Ecclesiastical History:* 55 (4): 726-737.

Buurma, Rachel Sagner, Anna Tione Levine, and Richard Li. 2011. The Early Novels Database: a case study. Available at: http://www.academiccommons.org/commons/essay/early-novels-database.

Clark, Mark, Gill Ferrell, and Paul Hopkins. 2011. Study of early adopters of shared services and cloud computing within Higher and Further Education. Available at: http://he-associates.co.uk/FEAST.aspx.

Cushing, Judith Bayard, Nalini Nadkarni, Lois Delcambre, Keri Healy, Dave Maier, and Erik Ordway. 2002. The development of databases and database tools for forest canopy researchers: a model for database enhancement in the ecological sciences, 2002. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.1176.

Cushing, Judith Bayard, Nalini Nadkarni, Michael Finch, Anne Fiala, Emerson Murphy-Hill, Lois Delcambre, and David Maier. 2007. Component-based end-user database design for ecologists. *Journal of Intelligent Information Systems:* 29 (1): 7-24.

Evrenosoglu, Faik Burak. 2008. Use of relational databases in forensic delay analysis. *AACE International Transactions*: 1-13.

Goth, Gregory. 2010. Turning data into knowledge. *Communications of the ACM*; 53(11): 13-15. Available at: http://portal.acm.org/citation.cfm?doid=1839676.1839682.

Hamilton, James. 2008. Interview: database dialogue with Pat Selinger. *Communications of the ACM*: 51(12): 32-35. Available at: http://portal.acm.org/citation.cfm?doid=1409360.1409373.

Hine, Christine. 2006. Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science*: 36(2): 269-298. Available at: http://sss.sagepub.com/cgi/doi/10.1177/0306312706054047.

Johnson, Todd M. 2005. Collaboration and the World Christian Database. *Missiology: An International Review* XXXIII (1).

Key Perspectives Ltd. 2010. *Data dimensions: disciplinary differences in research data sharing, reuse and long term viability.* A DCC-commissioned report, available at: http://www.dcc.ac.uk/sites/default/files/documents/publications/case-studies/SCARP SYNTHESIS.pdf.

OECD Follow Up Group. 2003. Promoting access to public research data for scientific, economic, and social development. Available at: http://dataaccess.ucsd.edu/Final_Report_2003.pdf.

Research Information Network. 2011. Reinventing research? Information practices in the humanities. Available at: http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/information-use-case-studies-humanities.

Seltzer, Margo. 2008. Beyond relational databases. *Communications of the ACM*: 51(7): 52-58. Available at: http://portal.acm.org/citation.cfm?doid=1364782.1364797.

Van Den Eynden, Veerle, Louise Corti, Matthew Woollard, Libby Bishop, and Laurence Horton. 2011. Managing and sharing data. Available at: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf.

Westra, Brian. 2010. Data services for the sciences: a needs assessment. *Ariadne*: 64. Available at: http://www.ariadne.ac.uk/issue64/westra/.

Willcocks, S., W. Venters, and E. Whitley. 2011. Cloud and the future of business: from cost to innovation. Available at: http://www.outsourcingunit.org/publications/cloudPromise.pdf.

Wilson, James A.J. and Meriel Patrick. 2010. *Sudamih Researcher Requirements Report*. Available at: http://sudamih.oucs.ox.ac.uk/docs/Sudamih Researcher Requirements Report.pdf.