

E-science in the Library and Information Studies Sector

Overview of seminar findings for the e-science scoping study

Melissa Terras 24/07/06

“We’re very depressed by the challenges”. Quote from the LIS seminar.

The field of Library and Information Science has a history of embracing technology, and aiming to understand how that will affect how we create, access, store, and retrieve data and information. Much of the research into information environments being undertaken nowadays involves some study of how technology is being used in the sector, as it becomes an increasingly technologically literate discipline. Nevertheless, there are challenges and barriers for the sector in the uptake and use of e-science. Some of these are practical: getting access to technologies, understanding how they may be used, and having the expertise and skills in order to use them to advance either service or research within the LIS sector. Some barriers and challenges are more fundamental: what part of LIS research can e-science contribute to? What are the tools it will provide to analyse data – but what data is there to analyse? What is the research question which can only be answered through the use of e-science technologies? If LIS tends to study information environments, and their successes and failures, how can they study the application and use of e-science technologies when it is a little known technology with few available applications to study how users are manipulating information sources? This overview of the discussion seminar on the potential in e-science for the LIS sector will cover some challenges and barriers for the uptake of the technology, highlight some potential applications of e-science for LIS, and give some recommendations for an e-science and the arts agenda.

1. Challenges and Barriers for the uptake of e-science in LIS

Like any other discipline, LIS needs to be inspired by e-science technologies to produce new research: however, without adequate explanation of e-science in a terminology which LIS professionals understand, and without adequate access to the technologies involved, LIS professionals will neither be able to formulate research questions which adopt and adapt e-science methodologies, nor have the ability to utilise technologies which may prove useful to them. There are issues about gaining access to the grid at the moment, and how to “buy” computational power – at the moment you may have to spend a long time queuing for the computational power – another barrier to access. These barriers are fundamental, in that no research in LIS can be undertaken unless they are overcome by individual researchers who may have research ideas which would require such technologies. Although institutional support can have an affect on providing the type of training and access to these technologies required, until individual researchers can apply these technologies, have the opportunity to access and experiment with them, or the opportunity to collaborate with others who have the expertise required to undertake e-science analysis, little research of use will be done with e-science in the LIS sector.

Arts and humanities dataset are often smaller, and more complex, than the type of datasets that allows data science methodologies to be utilised. Understanding the nature of humanities datasets is fundamental to understanding how e-science technologies can be applied to them. It is often the case that arts and humanities datasets are created with a representation and format that meets users immediate needs, and not a reusable amount or type of information that would be of interest to other scholars. Discovering datasets which are large enough to warrant e-science

processing is an issue for the arts and humanities as a whole. The LIS sector tends to have larger bibliographic datasets, but whether these would warrant analysis using e-science methodologies is questionable.

As arts and humanities users of computers, LIS professionals are often used to approaching computing programmes which are already fully formed and tested, and often commercially available. However, most work undertaken in the sciences and e-science is still at the stage where the technology is in development. Standards do not exist. Tools are not available which can be used on any other task than the project they were programmed for. The tools that do exist are not “useable” by users without in-depth technological understanding, as little work on usability or human computer interaction has been done on these interfaces and systems. Finding the technological understanding to develop our own tools, or entering the dialogue at this late stage in the building of the e-science infrastructure, whilst understanding the terminology and (scientific) domain specific nomenclature and dialogue which has taken place to construct that infrastructure, is a real barrier to adoption of e-science in the LIS sector, or the repurposing of tools from the scientific domain for LIS.

This will be a similar problem across the arts and humanities domains: presenting the technologies in a manner which is understandable and useful for the average humanities professional is a real challenge. Perhaps there is room to build better links between current work which has been undertaken in the sciences and social sciences with the arts and humanities – by aiding these domains to undertake outreach activities, arts and humanities subject areas such as LIS could be encouraged to uptake the technology.

It is important to realise that encouraging dialogue between these different domains is not a one way process. Librarians, and those trained in information science, have a lot to offer the scientific community who utilising e-science, grid and high performance computing technologies. Often scientific projects deal with vast datasets: the LIS community has a role in providing expertise regarding how this data may be stored, used, and accessed in the future. Without adequate and realistic documentation and storage mechanisms, the vast datasets being generated will be unusable to other projects, or in the future: therefore the complexity and size of datasets used in e-science type computing may preclude it from being used in future e-science type initiatives unless made available in standardised and documented format. Grid enabling should also include documentation, and LIS professionals have a lot to bring to the general e-science community about the practicalities of storing data efficiently. In turn, Librarians can learn different aspects of data management and manipulation through working with large scale datasets in the scientific domain, which may produce new research in LIS itself, in the development of a shared conceptual framework and a long-term strategy for a sustainability policy which makes it possible to reuse e-science datasets (and reformat them in order to reuse them).

2. Potential of e-science for LIS

One of the fundamental aspects of LIS research is to study use and user behaviour of information. E-science technologies, and the processing power of the grid, may be useful to share information about users (massive sets of server log data, for example) and be able to analyse these huge datasets effectively. E-science and LIS then becomes a technology which can be used to understand and profile user behaviour in

cyberspace: an increasingly important research topic, with few methodologies defined and a paucity of robust, statistically verifiable research available . It will be possible to analyse large datasets of user behaviour quickly and efficiently, and to share this data with others, in order to further understand the use of a popular yet complex environment, which has use for not only the LIS sector but the general online community too. The users of technology are the most important and often most forgotten about link in the technological chain: in this case, a very new technology (e-science) can be used to study a maturing technology (the internet): data mining can be used to study the use of data generated from digital environments themselves, on an unparalleled scale. Understanding user interaction with technology is an expensive and difficult task: Library and Information Professionals are often trained in how to study information environments (with both qualitative and quantitative methods). This would inform us about retrieval, and searching, and navigation: the fundamental things that information professionals are supposed to be all about.

The study of e-science environments would be another relevant and important research topic. There is concern about the replication of tools, the non-effectiveness of most tools, and the parochial nature of e-science tools which have been developed for individual projects which cannot be reused or developed for other projects. Common methodologies and systems will have to be developed to enable e-science tools and techniques to be rolled out over a variety of disciplines without the massive funding initiatives which we have seen to date (which are drying up). Investigating the e-science environment and the tools they use and utilise to find common characteristics, using well established LIS methodologies for the study of information environments, will bring further understanding of these issues to the wider academic community.

Further potentials lie in the development of text mining, data mining, and information retrieval tools that would utilise the processing power of the grid to aid the LIS professional. The delineations between these can be hard to define (and also who should be developing these: the LIS professional, computer scientist, or engineer? There is further room for LIS professionals to be involved in the development of tools in this arena.)

The ‘e- arts and humanities Agenda’

Three major items came out of a discussion regarding the wider application of e-science in the arts and humanities, as topics that should be addressed in an e-arts and humanities agenda. Firstly, there is a need to classify and organise information for discovery and reuse. Secondly, there is a need to understanding the collaboration of scholars and the collaborative use of datasets and tools. Thirdly, there is need for a set of tools to mine large-scale datasets, to provide tools for analysing large amounts of user behavioural data.

There is room for a shared infrastructure and policy on data management, storage, and formatting, for e-science projects, in order to create reusable datasets which are reliable, authentic, and have integrity. The licensing requirements on this also need to be investigated, and how this data can be disseminated. Arts and Humanities scholars, as a community, should also be aware that commercial firms (such as the Googles of this world) are very interesting in acquiring large datasets and repurposing them for their own means, and as a field we should be aware that this cultural shift from the lone scholar to the cyber community is coming. It is important that we

understand the ramifications of data use and reuse in this environment (as well as patterns of behaviour of users in this environment).

There is a role for collaborative technology to aid in the use and reuse of datasets. Recommender systems to datasets (a relatively cheaply available collaboration technology) could be used to encourage reuse. A framework needs to be put in place to aid scholars in collaborating with e-science technologies, and to develop tools which can be repurposed (or repurpose previously developed tools) to cut down on intellectual duplication and wastage.

Sets of tools which are available to mine large datasets are required. There is a need to focus on extraction retrieval issues, in order for us to query large datasets, in the way we want to, effectively. More attention should be paid to finding, choosing, using and manipulating available data.

There was some discussion of specific tools and applications – such as a strategic lightweight annotation tool for creating interoperable grid enabled datasets, incorporating textual markup and metadata annotation. There was also talk of building a grid capable inference engine which could run an analysis project on all the metadata created by all the eminent libraries: utilising the processing power of the grid to do advanced entity extraction, utilising ontologies, to aid data finding and extraction. Building tools to aid interpretation, and interpretation sharing, tools which could aid in the conversion of data, and tools which could aid in information retrieval would encourage the use of e-science and related technologies across all arts and humanities subject areas.

Additionally, the construction of a reservoir of user data with the various kinds of tools for analysis presented, to analyse the large log type datasets which are emerging from various online environments would be useful to researchers. It is the first time that evidence of this nature and scale has become available regarding how users interact with online environments, and LIS and the arts and humanities should be embracing this.

Finally, these activities should not be exclusive of each other. It is possible to imagine an information environment where we have user behaviour, and where we can actually look at what people are using as they are searching through the data. It will be possible to cluster these and build up a picture of collaborative data searching and sharing, and an overview of people using the same techniques, and the results of their searching. It will be possible to look at techniques in data and text mining which mimic those searches, in order to compare the weaknesses and strengths of the various approaches to data manipulation. This combination of user behaviours, collaborative environments and data mining may allow us to investigate, query, and ultimately understand what e-science technologies bring to the arts and humanities, and the role they play in our understanding of data and techniques to manipulate and repurpose it.