

## **IT Challenges for the Library and Information Studies Sector**

This document is intended to facilitate and stimulate discussion at the e-Science Scoping Study Expert Seminar for Library and Information Studies. Essentially, it is a tentative specification and delineation of challenges and solutions for IT in the Library and Information Studies sector. It will be of particular use in the morning session, and during our discussion on ‘Challenges for Library and Information Studies in the Information Age’.

It begins by suggesting six important areas where IT is used in the Library and Information Studies Sector, before looking in more detail at each one. It suggests some specific barriers that apply, and, in some cases, potential e-science solutions to them. We plan to discuss each area in turn, with a particular emphasis on specifying the pressing issues, challenges and barriers that apply, and suggesting how developments in technology can (or might) aid in their solution. Please feel free to disagree with this document, or suggest other areas where IT is used in the LIS sector. As experts in LIS, one of the most important contributions you can make to the eSSS is to ascertain the main Information Technology challenges for the LIS sector.

Important areas where IT is used in Library and Information Studies might be defined as follows

- 1.) Content creation and digitization
- 2.) Information and Knowledge Management
- 3.) Information Retrieval
- 4.) Legal and copyright issues, including access, and authentication
- 5.) Processing of large volumes of data, including cross collection processing
- 6.) Understanding user requirements, and building usable and accessible interfaces to large datasets, and e-science technologies.

### **1. Content creation and digitization**

## **Summary**

New digital information is being generated all the time in all academic areas. This information needs to be properly stored, and managed, in order for it to be as searchable as possible- to ensure it remains reliable and easy to find. Both born digital and digitized material needs to be managed and curated properly to facilitate searching, analysis, and preservation of the material. Challenges exist to ensure cross platform standards are developed and maintained for digital data to ensure its longevity and usefulness.

There is also an increased public demand for digitized library materials, and the act of digitization is becoming more common. Creating digitized records, texts, images, and other media is a costly and time-consuming endeavour. Projects such as GooglePrint are in the process of trying to digitise as much as possible as quickly as possible- enabling a private user to search the pages of millions of uncategorized, unmarked-up and untagged digitized books. While a formidable (if legally controversial) resource, there is also concern that the quality of the searches from Googleprint are often low. Overall, the cheap mass-digitization from Googleprint is creating a deluge of relatively poorly described data. While there is nothing wrong with what Googleprint is doing as such, it is necessary that the texts that have been digitized are marked up and tagged universally and appropriately or that there are developments in data mining undertaken to allow users to search through vast amounts of data more intuitively.

## **Specific Barriers**

- -The Development of content creation standards
- -Maintenance of material
- -Management of material
- -Sheer volume of data to manage
- -Volume of data to utilise and integrate

It can be assumed that the majority of the vast volumes of information being created in digital format will not be provided with adequate metadata or marked up in any way to facilitate cataloguing or retrieval. Therefore technologies will need to be developed to aid in intelligent sorting of such material. Standards will still be required from the LIS sector to provide metadata for information – however, is this leading to a two tier information environment; a voluminous free for all of uncatalogued data, and a small carefully provided set of catalogued material? How can these be integrated? How can the wider audience be encouraged to maintain and catalogue their data? Is this an issue, or should LIS professionals be concentrating on safeguarding information from their own institutions, and providing a small subset of data of high quality

### **How these might be overcome by e-science?**

Maybe the question is: how can e-science deal with these problems? E-science is just another amorphous buzzword for a set of already available and established technologies. Using advanced technologies will not make these problems go away – if anything, it will bring them to the fore. We need to understand what to do with them when they arise, but e-science will not and cannot stop the flow of digital information.

E-science can certainly provide the processing power to sort through vast digitised datasets –however, more intelligent searching algorithms are needed and data mining techniques required to ensure that these datasets will be useful. What techniques will be required by the LIS sector?

## **2. Information and Knowledge Management**

### **Summary**

Metadata needs to be standardised and universalised if it is to work properly and in harmony. The lack of any agreed metadata standard in the classification of images, for example, is a real problem. Essentially, standards need to be agreed upon and implemented to solve the problem.

The same is true of semantic web ontologies. The semantic web has great potential, but will rely on the agreement of standards by subject experts- for example what in history is a primary source, and what is a secondary source? The semantic web may be too difficult to understand- certainly more so than html- which limits its potential. The automated production of ontologies using the processing power of the grid may be a future development which would aid the LIS sector: but relevant tools and technologies have yet to be implemented which would do this in a trustworthy and useful manner.

There is also the evolution of organic metadata in the form of 'Folksonomies' to consider. These organic 'ontologies of how the world works' are very different from formal 'top down' Metadata methodologies. Created by users, rather than professionals, folksonomies are the network of interconnected terms used to describe data, such as the picture content in Flickr, and are chosen entirely by the general public, developing a structure through volume of entries received. Naturally, this might theoretically be a great asset to the production of metadata vocabularies and taxonomies, but folksonomies have their drawbacks. They sometimes apply only for a limited time- for example 'Danish' and 'Cartoons' have become related words in response to developments in the news- but in time these words will have lost their association. Sometimes they can also be judgemental- eg, 'Bush' and 'idiot' are related words in the Googlesphere. Thus folksonomies suit free-searching and must be treated with care if given any official support by library and information scientists.

### **Specific Barriers?**

There are already many systems, standards and recommendations in place for the provision of metadata. How can LIS professionals cope with the amount of information produced which needs cataloguing? How can the general user be encouraged to learn about these standards? Should these standards be changed to encompass e-science technologies to facilitate its uptake? How can the LIS professional learn what is best for using e-science? What can the LIS professional give back to e-science regarding the management and organisation of large scale datasets?

### **How these might be overcome by e-science?**

How can e-science be used to produce metadata? What can the LIS professional bring to the e-science agenda in terms of how large volumes of data should be stored and managed?

## ***3. Information Retrieval***

### **Summary**

The sheer amount of data which currently exists poses problems for the LIS sector. There is an increasing expectation that Libraries must have complete searchable catalogues of all they hold- which at the moment is becoming more common but is not universally implemented. The nature of archival material, for example, means that it will rarely be exhaustively catalogued.

The demand has been exacerbated by perceived use of Google as a library. Libraries need time to organise their existing holdings with appropriate metadata to store things logically to create holdings of organised, searchable, and reliable data.

Although Google may represent a challenge, the quality of the information it supplies is low, and it is still often quicker to find more reliable information through libraries. (This has also been shown in recent studies/ experiments).

## **Specific Barriers?**

How can LIS professionals deal with the large amount of digitised information available? Should this be seen as separate to the work of LIS? How can their information retrieval tools and techniques be used in the broader environment, or how can they be developed further to deal with vast amounts of data?

## **How these might be overcome by e-science?**

IR is the area where e-Science may have its biggest successes in the LIS sector: given the massive processing power available through the grid, it should be possible to deal with larger volumes of data, and process elaborate queries in a realistic time frame. However, this requires understanding of the technologies themselves, and perhaps without integration with computer scientists, complex IR developments for LIS will not become available. How can the library world liaise more closely with those developing and utilising e-Science technologies?

## ***4. Legal Challenges (copyright issues, access, authentication, reliability)***

### **Summary**

Individuals, through the use of privately owned hardware and the internet, (and to an extent Google) are gnawing away at the Library's previous relative monopoly on published and archival material. Although this leads to greatly enhanced access to materials, the quality (for reasons discussed above), and legality of such amateur material is questionable. Googleprint is involved in multiple lawsuits at present on whether it can allow texts, or portions of text, to be viewable online. It is true that many journal articles and texts are already available online, but these are secured through systems such as Athens. The academic community's faith in the Library is not in question, but Library and Information scientists need to find a way of maintaining a high quality, consistent, methodical, and vetted body of online

information which is secure and with limited access, or providing guidance in the creation of “amateur material” and an understanding of how this may be used (through technologies such as e-Science to provide effective and complex searches and retrieval of information).

The LIS sector also has to ensure that their own data abides with copyright law and that the information they provide is authentic, reliable, and accessible.

### **Specific Barriers?**

Legal issues must always be taken into consideration when providing digitised content.

### **How these might be overcome by e-science?**

E-science cannot overcome the manmade legal system. However, how can standards be developed to allow licenses to be integrated into large scale systems?

## ***5. Processing of large volumes of data, including cross collection processing***

### **Summary**

Digital information comes in many different formats, and is often stored and managed by bespoke systems in different institutions. Cross collection searching and processing can then be hampered by differing formats and access issues. The processing of large volumes of data often means for that data to be converted into like for like formats, allowing for fair comparison and easier information retrieval.

### **Specific Barriers?**

Conversion of materials from different formats and systems is time consuming, costly, and can often be technically demanding.

**How these might be overcome by e-science?**

Again, e-science may aid in this task – the large processing power available would allow the conversion of large volumes of data quickly, and aid in the complexities of cross collection searching and processing.

***6 Understanding user requirements, and building usable and accessible interfaces to large datasets, and e-science technologies***

The uptake of all technology is dependent on the technologies matching user requirements and those technologies being easily implemented, managed, and used. Although those in the LIS sector have a growing understanding of IT, e-science technologies remain obscure: both in its function, and their availability. How can e-science systems be developed which look seamless and are intuitive for both LIS professionals and users? How can e-science become more available?