e-Science and its implications for the library community

Tony Hey
Microsoft Corporation, Redmond, USA
Jessie Hey
School of Electronics and Computer Science and University of Southampton Libraries, University of Southampton, Southampton, UK

Abstract

Purpose: To explain the nature of the 'e-Science' revolution in 21st century scientific research and its consequences for the library community.

Design/methodology/approach: The concepts of e-Science are illustrated by a discussion of the CombeChem, eBank and SmartTea projects. The issue of open access is then discussed with reference to arXiv, PubMed Central and EPrints. The challenges these trends present to the library community are discussed in the context of the TARDis project and the University of Southampton Research Repository.

Findings: Increasingly academics will need to collaborate in multidisciplinary teams distributed across several sites in order to address the next generation of scientific problems. In addition, new high-throughput devices, high resolution surveys and sensor networks will result in an increase in scientific data collected by several orders of magnitude. To analyze, federate and mine this data will require collaboration between scientists and computer scientists; to organize, curate and preserve this data will require collaboration between scientists and librarians. A vital part of the developing research infrastructure will be digital repositories containing both publications and data.

Originality/value: The paper provides a synthesis of e-Science concepts, the question of open access to the results of scientific research, and a changing attitude towards academic publishing and communication. The paper offers a new perspective on coming demands on the library and is of special interest to librarians with strategic tasks.

Keywords: Digital libraries, Digital storage

Paper type: Research paper

Introduction

As Thomas Friedman (2005) eloquently explains in his book 'The World is Flat', the convergence of communication and computing technologies is changing the world of both business and leisure. It would be naïve to think that the academic research community will be immune from these changes. The methodology of research in many fields is changing and we are on the threshold of a new era of data-driven science. In the last few decades computational science has emerged as a new methodology for scientific research on an equal footing with the traditional experimental and theoretical methodologies. Simulation is now used as a standard weapon in the armoury of the scientist to explore domains otherwise inaccessible to the traditional research methodologies - such as the evolution of the early universe, the design of new materials, the exploration of climatology over geological timescales and, of course, the weather forecasts we now take for granted. Its use in industry is becoming even more widespread with computational fluid dynamics and finite

element simulations now an essential part of the design process, complementing traditional experimental wind tunnel and safety testing in the aero and auto manufacturing industries, with simulations of oil fields and analysis of seismic data now playing a key role in the oil and gas industry, and with simulation playing an increasingly important role in the drug design life cycle in the pharmaceutical industry.

The next decade will see the emergence of a new, fourth research methodology, namely 'e-Science' or networked, data-driven science. Many areas of science are about to be transformed by the availability of vast amounts of new scientific data that can potentially provide insights at a level of detail never before envisaged. However, this new data dominant era brings new challenges for the scientists and they will need the skills and technologies both of computer scientists and of the library community to manage, search and curate these new data resources. Libraries will not be immune from change in this new world of research. The advent of the Web is changing the face of scholarly publishing and the role of publishers and libraries. The National Science Foundation Blue Ribbon Report on Cyberinfrastructure lays out a vision of this new world. On publishing, the report states:

> The primary access to the latest findings in a growing number of fields is through the Web, then through classic preprints and conferences, and lastly through refereed archival papers. (Atkins *et al*., 2003, p. 9)

And on scientific data the report states:

> Archives containing hundreds or thousands of terabytes of data will be affordable and necessary for archiving scientific and engineering information. (Atkins *et al*., 2003, p. 11)

This paper explores some of the challenges facing both the scientific and library communities in this new emerging world of research and delineates the key role that can be played by computer science and by IT companies such as Microsoft in assisting the research community.

e-Science and Licklider's vision

It is no coincidence that it was at CERN, the particle physics accelerator laboratory in Geneva, that Tim Berners-Lee invented the World Wide Web. Given the distributed nature of the multi-institute collaborations required for modern particle physics experiments, the particle physics community urgently needed a tool for exchanging information. It was their community who first enthusiastically embraced the Web as a mechanism for information exchange within their experimental collaborations and it was no accident that the first Web site in the USA was at the Stanford Linear Accelerator Center Library. As we all now know, since its beginnings in the early 1990's, the Web has not only taken the entire scientific world by storm but also the worlds of business and leisure. Now, just a decade or so later, scientists need to develop capabilities for collaboration that go far beyond those of the original World Wide Web. In addition to being able just to access information from different sites, scientists now want to be able to use remote computing resources, to integrate, federate and analyze information from many disparate and distributed data resources, and to access and control remote experimental equipment. The ability to access, move, manipulate and mine data is the central requirement of these new collaborative science applications – whether the data is held in flat files or databases, or is data

generated by accelerator or telescopes, or data gathered in real-time from potentially mobile sensor networks.

In the United Kingdom, at the end of the 1990's, John Taylor became Director General of Research Councils at the Office of Science and Technology (OST) in the UK – roughly equivalent to Director of the National Science Foundation (NSF) in the USA. Taylor had been Director of Hewlett-Packard (HP) Laboratories in Europe and HP's vision for the future of computing has long been that IT resources will become a new 'utility'. Rather than purchase IT infrastructure, users will pay for IT services as they consume them, in the same way as the conventional utilities such as electricity, gas and water – and now mobile telephones. In his role at the OST as overseeing the funding of UK scientific research, Taylor realized that many areas of science could benefit from a common IT infrastructure to support multi-disciplinary and distributed collaborations. He articulated a vision for this type of distributed, collaborative science and introduced the term 'e-Science':

> e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it. (Taylor, 2001)

It is important to emphasize that e-Science is not a new scientific discipline in its own right: e-Science is shorthand for the set of tools and technologies required to support collaborative, networked science. The entire e-Science infrastructure is intended to empower scientists to do their research in faster, better and different ways.

Of course, these problems are not new – the computer science community has been grappling with the challenges of distributed computing for decades. Indeed, such an e-Science infrastructure was very close to the vision that J.C.R. Licklider ('Lick') took with him to ARPA (Advanced Research Projects Agency) when he initiated the core set of research projects that led to the creation of the ARPANET. Larry Roberts, one of his successors at ARPA and principal architect of the ARPANET, described this vision as follows:

> Lick had this concept of the intergalactic network which he believed was everybody could use computers anywhere and get at data anywhere in the world. He didn't envision the number of computers we have today by any means, but he had the same concept – all of the stuff linked together throughout the world, that you can use a remote computer, get data from a remote computer, or use lots of computers in your job. The vision was really Lick's originally. (Segaller, 1998, p. 40)

The ARPANET of course led to the present day Internet - but the killer applications have so far been email and the Web rather than the distributed computing vision originally described by Licklider. Of course, in the early 1960's, Licklider was only envisaging connecting a small number of rather scarce and expensive computers, and at relatively few sites. However, over the past thirty years, Moore's Law – Gordon Moore's prediction that the number of transistors on a chip would double about every 18 months so that the price-performance is halved at the same time – has led to an explosion in the number of supercomputers, mainframes, workstations, personal computers and PDAs that are now connected to the Internet. Already we are beginning to see programmable sensors and RFIDs – intelligent tagging devices - being connected to the network.

An example of e-Science: The CombeChem, eBank and SmartTea projects

The CombeChem project [1] was funded by the Engineering and Physical Sciences Research Council in the UK and its goals were to enhance the correlation and prediction of chemical structures and properties by using technologies for automation, semantics and Grid computing (see Frey *et al.*, 2003; Hughes *et al.*, 2004). A key driver for the project was the fact that large volumes of new chemical data are being created by new high throughput technologies. One example uses the technologies of combinatorial chemistry in which large numbers of new chemical compounds are synthesized simultaneously. The volume of data and the speed by which it can be produced highlights the need for assistance in organizing, annotating and searching this data. The CombeChem team consisted of a collection of scientists from several disciplines – chemistry, computer science and mathematics – who developed a prototype test-bed that integrated chemical structure-property data resources with a 'Grid' style distributed computing environment. The project explored automated procedures for finding similarities in solid-state crystal structures across families of compounds and evaluated new statistical design concepts in order to improve the efficiency of combinatorial experiments in the search for new enzymes and pharmaceutical salts for improved drug delivery.

The CombeChem project also explored some other important e-Science themes. One theme concerned the use of a remote X-ray crystallography service for determining the structure of new compounds. This service can be combined in workflows with services for computer simulations on clusters or searches through existing chemical databases. Another important e-Science theme was the exploration of new forms of electronic publication – both of the data and research papers. This e-Publication theme was examined in the eBank project [2] funded by the Joint Information Systems Committee (JISC). One of the key concepts of the CombeChem project was that of 'Publication@Source' which establishes a complete end-to-end connection between the results obtained at the laboratory bench and the final published analyses (Frey *et al.*, 2002). This theme is linked to yet another of the e-Science themes explored in the CombeChem project that was concerned with human-computer interfaces and the digital capture of information. In the associated SmartTea project [3], computer scientists studied the way chemists within the laboratory used their lab notebooks and developed acceptable interfaces to handheld tablet technology (see Schraefel *et al.*, 2004a; Schraefel *et al.*, 2004b). This is important since it facilitates information capture at the very earliest stage of the experiment. Using tablet PCs, the SmartTea system has been successfully trialed in a synthetic organic chemistry laboratory and linked to a flexible back-end storage system. A key usability finding was, not surprisingly, that users needed to feel in control of the technology and that a successful interface must be adapted to their preferred way of working. This necessitated a high degree of flexibility in the design of the lab book user interface. The computer scientists on the team also investigated the representation and storage of human-scale experiment metadata and introduced an ontology to describe the record of an experiment.

A novel storage system for the data from the electronic lab book was also developed in the project. In the same way that the interfaces needed to be flexible to cope with whatever chemists wished to record, the back end solutions also needed to

be similarly flexible to store any metadata that might be created. This electronic lab book data feeds directly into the scientific data processing. All usage of the data through the chain of processing is now effectively an annotation upon it, and the data provenance is explicit. The creation of original data is accompanied by information about the experimental conditions in which it is created. There then follows a chain of processing such as aggregation of experimental data, selection of a particular data subset, statistical analysis and modeling and simulation. The handling of this information may include explicit annotation of a diagram or editing of a digital image. All of this generates secondary data, accompanied by the information that describes the process that produced it. This digital record is therefore enriched and interlinked by a variety of annotations such as relevant sensor data, usage records or explicit interactions. By making these annotations machine processable, they can be used both for their anticipated purpose and for subsequent unanticipated reuse. In the CombeChem project this was achieved by deployment of Web Services and Semantic Web technologies (Berners-Lee *et al.*, 2001). RDF (Resource Description Framework) was used throughout the system: at present there are over 70 million RDF triples in the CombeChem triplestore. This system was found to give a much higher degree of flexibility to the type of metadata that can be stored compared to traditional relational databases.

In the sister eBank project, raw crystallographic data was annotated with metadata and 'published' by being archived in the UK National Data Store as a 'Crystallographic e-Print' [2]. Publications can then be linked back directly to the raw data for other researchers to access and analyze or verify. Another noteworthy feature of the project was that pervasive computing devices were used to capture laboratory conditions so that chemists could be notified in real time about the progress of their experiment using hand held PDAs.

The imminent data deluge: A key driver for e-Science

One of the key drivers underpinning the e-Science movement is the imminent availability of large amounts of data arising from the new generations of scientific experiments and surveys (Hey and Trefethen, 2003). New high-throughput experimental devices are now being deployed in many fields of science - from astronomy to biology - and this will lead to a veritable deluge of scientific data over the next 5 years or so. In order to exploit and explore the many Petabytes of scientific data that will arise from such next-generation scientific experiments, from supercomputer simulations, from sensor networks and from satellite surveys, scientists will need the assistance of specialized search engines and powerful data mining tools. To create such tools, the primary data will need to be annotated with relevant metadata giving such information as to the provenance, content and the conditions that produced the data. Over the course of the next few years, scientists will create vast distributed digital repositories of scientific data that will require management services similar to those of more conventional digital libraries as well as other data-specific services. As we have stressed, the ability to search, access, move, manipulate and mine such data will be a central requirement – or a competitive advantage - for this new generation of collaborative data-centric e-Science applications.

With this imminent deluge of scientific data, the issue of how scientists can manage these vast datasets becomes of paramount importance. Up to now, scientists have generally been able to manually manage the process of examining the experimental data to identify potentially interesting features and discover significant relationships between them. In the future, when we consider the massive amounts of data being created by simulations, experiments and sensors, it is clear that in many fields they will no longer have this luxury. The discovery process - from data to information to knowledge – needs to be automated as far as possible. At the lowest level, this requires automation of data management with the storage and organization of digital entities. At the next level, we require automatic annotation of scientific data with metadata describing both interesting features of the data and of the storage and organization of the resulting information. Finally, we will need new tools to enable scientists to progress beyond the generation of mere structured information towards the automated knowledge management of our scientific data.

The future of scholarly communication

The Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities was drafted in 2003 'to promote the Internet as a functional instrument for a global scientific knowledge base and human reflection and to specify measures which research policy makers, research institutions, funding agencies, libraries, archives and museums need to consider' (Berlin Declaration, 2003). Signatories to the original declaration included research organizations such as the Fraunhofer and Max Planck Institutes in Germany, Centre National de la Recherche Scientifique (CNRS) and Institut National de Recherche en Informatique et en Automatique (INRIA) in France, Royal Netherlands Academy of Arts and Sciences (KNAW) and SURF in the Netherlands, JISC in the UK, CERN and Swiss Federal Institute of Technology (ETH) in Switzerland as well as many other international organizations and universities. The Berlin meeting followed in the footsteps of the Budapest Open Access Initiative in 2001 [4]. It is important to recognize that the Berlin Declaration is not just concerned with textual material. The declaration defines open access contributions to include "original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material" (Berlin Declaration, 2003).

The research community is responding to the challenge of open access in a number of ways. Consider the three leading 'prophets' of open access – Paul Ginsparg of arXiv [5], David Lipman of PubMed Central [6] and Stevan Harnad of EPrints [7].

The theoretical particle physics community had long had a tradition of circulating hard copy preprints of papers submitted to conventional journals ahead of review and publication. In the fast moving field of theoretical physics, the community is used to discussing the latest ideas at informal seminars and workshops and it makes no sense to attempt to delay discussion until after formal publication. With such a well established preprint tradition, it was a natural but very significant step for Paul Ginsparg to establish an electronic archive at Los Alamos, where e-prints, electronic versions of preprints, could be displayed on a web site. From these small beginnings, Ginsparg has demonstrated a new way of scholarly communication

outside the traditional scholarly publishing route of refereed journal articles. The arXiv has now moved to Cornell where it is owned and managed by the Cornell library and this is now the standard first port of call for scientists in several subfields of physics, mathematics, computer science and quantitative biology [5]. It is interesting that the arXiv has no formal refereeing process to restrict publication on the site. Perhaps it is the very mathematical nature of the field that prevents the site from being overwhelmed by 'noise' of low quality material. This mode of publication leads to many headaches for librarians of course. The published journal version of the original e-print may have revisions to the text and will certainly have a different layout and pagination. Proliferation of versions – e-prints, preprints, postprints and so on – as well as confusion about the precise date of 'publication' are all now areas of concern to librarians. From a scientific point of view, these issues may seem trivial - since there is no doubt that claims for priority would be determined by the date of the e-print – but they are not at all trivial from the perspective of librarians and archivists [8].

The National Institutes of Health (NIH) in the USA has a mandate to make publicly available a National Library of Medicine (NLM) of biomedical and healthcare resources. Their Entrez Life Sciences Search Engine gives access to both the PubMed service containing over 16M citations from the MEDLINE database and life science journals for biomedical articles going back to the 1950's as well as a wide collection of biological databases. In February 2005, the NIH announced a new policy designed to accelerate the public's access to published articles resulting from NIH-funded research. The policy calls on scientists to release to the public manuscripts from research supported by NIH as soon as possible, and within 12 months of final publication. These peer-reviewed, NIH-funded research publications are now available in PubMed Central (PMC), a Web-based archive managed by the National Center for Biotechnology Information (NCBI) for the NLM [6]. The online archive will increase the public's access to health-related publications at a time when demand for such information is on a steady rise. In their announcement, NIH Director Elias A. Zerhouni, M.D. said:

> With the rapid growth in the public's use of the Internet, NIH must take a leadership role in making available to the public the research that we support. While this new policy is voluntary, we are strongly encouraging all NIH-supported researchers to release their published manuscripts as soon as possible for the benefit of the public. Scientists have a right to see the results of their work disseminated as quickly and broadly as possible, and NIH is committed to helping our scientists exercise this right. We urge publishers to work closely with authors in implementing this policy. (NIH News, 3 February 2005)

The NIH policy for PubMed Central has several important goals, including:

- Creating a stable archive of peer-reviewed research publications resulting from NIH-funded studies to ensure the permanent preservation of these vital research findings;
- Securing a searchable compendium of these research publications that NIH and its awardees can use to manage more efficiently and to understand better their research portfolios, monitor scientific productivity, and, ultimately, help set research priorities; and
- Making published results of NIH-funded research more readily accessible to the public, health care providers, educators, and scientists.

Beginning May 2, 2005, the policy requests that NIH-funded scientists submit an electronic version of the author's final manuscript, upon acceptance for publication, resulting from research supported in whole or in part by NIH. The author's final manuscript is defined as the final version accepted for journal publication, and includes all modifications from the publishing peer review process. The present policy gives authors the flexibility to designate a specific time frame for public release — ranging from immediate public access after final publication to a 12 month delay — when they submit their manuscripts to NIH. Authors are strongly encouraged to exercise their right to specify that their articles will be publicly available through PMC as soon as possible. With the addition of PubMed Central, Entrez searches can now be directed to free full text versions of the research article.

Jim Gray and Jean Paoli from Microsoft have worked with David Lipman and the NCBI team to develop a 'portable' version of PubMed Central which is now being deployed in other countries around the world. The NLM's archiving template for XML documents – the Document Type Definition or DTD – is now becoming the international standard for such archives. The Wellcome Trust in the UK, in partnership with the JISC and the NLM are working together on a project to digitise the complete backfiles of a number of important and historically significant medical journals [9]. The digitized content will be made freely available on the Internet – via PMC – and augment the content already available there. The Wellcome Library exists as a resource to provide access to the documentary record of medicine. This project is one way of translating that vision into the digital age.

The two repositories described above are examples of subject specific repositories. By contrast, Stevan Harnad advocates author 'self-archiving' in departmental or institutional repositories (Harnad and Hey, 1995). Open Access Archives or Repositories are digital collections of research articles that have been placed there by their authors. In the case of journal articles this may be done either before (preprints) or after publication (postprints). These repositories expose the metadata of each article (the title, authors, and other bibliographic details) in a format compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [10]. As a result, OAI-compliant search engines can harvest the metadata from each repository into large databases of worldwide research, which researchers can then use to locate articles of interest. Open access repositories can be centralised and subject-based, such as arXiv and PubMed Central, or they may be distributed and multidisciplinary, located in universities or other research-based institutions. A list of Open Access Archives is maintained at the Registry of Open Access Repositories (ROAR) [11] and OpenDOAR sites [12].

From these examples, one sees that the nature of scholarly publishing is changing. Not only is publication on the Web, in one form or other, enabling access to a much wider range of research literature but also we are seeing the emergence of data archives as a complementary form of scholarly communication. In some fields, such as biology, databases are already one of the primary mechanisms of scholarly publishing. In the area of environmental science, the NERC (Natural Environment Research Council, UK) DataGrid project is building a Grid that makes data discovery, delivery and use much easier than it is now, facilitating better use of the existing investment in the curation and maintenance of quality data archives [13]. This

DataGrid project also intends to make the connection between data held in managed archives and data held by individual research groups seamless in such a way that the same tools can be used to compare and manipulate data from both sources. When fully functional, it will deliver scientists the completely new ability of being able to compare and contrast data from an extensive range of US and European datasets from within one specific context. What is the role of the library community in this new world of scientific information management? One relevant example is that of the SPIRES Digital Library [14]. The SPIRES-HEP database has been run by the Stanford Linear Accelerator Center (SLAC) since the late 1960's as a database of particle physics literature. As mentioned in the introduction, this became the first web-site in North America in 1991 and now attracts around 50,000 searches per day from particle physicists. In contrast to just accessing research literature via arXiv, SPIRES offers access to the whole of the HEP literature with arXiv as only one of its key resources. The database is managed and maintained by the SLAC Library, in cooperation with librarians and particle physicists from DESY, FNAL, Kyoto, Durham, IHEP and KEK.

The above examples concern national and international data archives. However there is also likely to be a role for libraries at the institutional level in curating and preserving e-Science data in addition to their more traditional role in organizing and curating digital research output in the form of journal papers, reports and theses. Consider the following quotation from the rationale for MIT's DSpace repository:

> Much of the material produced by faculty, such as datasets, experimental results and rich media data as well as more conventional document-based material (e.g. articles and reports) is housed on an individual's hard drive or department Web server. Such material is often lost forever as faculty and departments change over time. (Tansley *et al*., 2003, p. 87)

Since some of this data may be relevant for the protection of the university's intellectual property, it is obvious that universities and libraries need to be clear about their roles in the curation and preservation of such data. In the next section, we describe a case study from the University of Southampton in the UK where there is experimentation with some of the different roles for the library in supporting research at the university.

An institutional repository at Southampton: The TARDis experience

Traditionally, academic libraries have played a major role in undergraduate teaching and information retrieval skills. The large increases in student numbers and pace of change in the Web based environment have caused librarians to constantly reexamine their skills and tasks for their support of undergraduate education. By contrast, research support has usually been concentrated on providing access to research resources external to the university and assisting the researcher in accessing original sources, e.g. by access to electronic versions of journals or to hard copy versions of papers and books via inter library loans. However, the growing emphasis on capturing research outputs at an earlier stage in the scholarly communication cycle now provides a significant opportunity for librarians to rethink their role as information managers, strategists and custodians of the research environment.

The JISC funded Focus on Access to Institutional Resources (FAIR) programme in the UK gave an opportunity to the Targeting Academic Research for Deposit and Disclosure (TARDis) project at the University of Southampton to investigate the practical implications of creating an Institutional Repository for research and to explore the relationship between an Institutional Repository and an Open Access Archive (Simpson and Hey, 2005).The project undertook a survey of the attitudes of the university's researchers - from senior management to individual academics – and their key feedback was to underline the importance of integrating a repository both into the university's current research management needs and also integrating the deposit process into the researcher's work practice. In the case of Southampton, a crucial functionality of the institutional repository was the ability to record publications for use by the university, by the department, by the research group and by individuals at an early stage in the scholarly research cycle - rather than at some more remote time such as that corresponding to formal publication which can be long after the initial production of a research output. Information capture could therefore take place either at the working paper stage or at the more final published paper and book chapter stage. Such a repository must empower academics to manage their own information management demands more efficiently and enable them to immediately add the full text or have the option to add this later if this is a more practical solution. A good summary of the goals of an Institutional Repository has been given by Lynch (Lynch, 2003).

The TARDis project was also able to feed information management requirements to the developers of the EPrints software at Southampton. In particular, they were able to influence the provision of fields and citation styles necessary to allow flexible reuse of the metadata. For example, in an institutional context, setting up a separate database solely for papers available with full text would require a huge duplication of effort if implemented on a university scale. The TARDis model therefore simply requires that searches of the whole database should reflect all types of research output and that searches for 'full text only' items can be obtained from the 'Open Access Archive' – the subset of research outputs for which the full text is stored on the same server. In the future, with changing attitudes to open access globally and with researchers becoming more familiar with saving and depositing their full text, the TARDis route map shows that the Institutional Repository comes closer to the vision of the Open Access archive (Hey *et al.*, 2005). However, with the increase in content in repositories such as arXiv and PubMed Central and other subject or conference based archives it seems increasingly likely that the research repository, as it grows in size and complexity, will be a pragmatic mix of full text, where the process of deposition is either straightforward or where there is a need to ensure there is a local copy, and of links to trusted repositories where this is more practical. By this time, it will be more meaningful for researchers to search the whole Institutional Repository rather than just the subsection of the archive which stores the full text locally. These are the kinds of information management decisions that librarians will have to make in the future just as they have traditionally weighed up whether to buy or acquire items on interlibrary loan. This mixed economy for the Institutional Repository has analogies with both the SPIRES database and its

relationship with arXiv, other full text repositories and journal sites and with the Entrez Search Engine, PubMed and the free full text PubMed Central repository.

The university publications database must portray the full picture of all research outputs: this is key to the goal of representing all disciplines fairly – not just those that follow the traditional peer reviewed scientific journal model. Listening to feedback from all parts of the academic community at Southampton has therefore resulted in a more complex project than had originally been envisioned by the library – namely, that of creating a full publications database rather than just a digital repository. However, this enlargement of scope has enabled the project to move from being just a pilot project to one that is now seen as an integral part of the university's research management infrastructure – and one that is able to respond to demands for Open Access on a more gradual but more sustainable timescale.

Where a national publications recording system is already in place such as in the Netherlands or in Australia, other information management decisions may need to be made. For example, there may be a need for practical steps to be taken by librarians to simplify recording in both a publications database and a full text database in an efficient manner (Woodland and Ng, 2006). In either case it is important for the original author to be part of the process to ensure the full text is deposited where possible and for the author to be aware of the potential for easy reuse of the metadata in CVs, project reports and proposals and for numerous other publication management demands.

The TARDis project was focused on research output but it is possible to envision a more ambitious role for an Institutional Repository as that of embracing the entire intellectual output of an institution. In working towards such a goal, there is much that the library can learn from the infrastructure required for a research library – both in recording research outputs and in the management of both publications and data. For example, the National Oceanography Centre at Southampton (NOC,S) is one of the world's leading centres for research and education in marine and earth sciences, for the development of marine technology and for the provision of large scale infrastructure and support for the marine research community. The National Oceanographic Library at NOC,S has long had a traditional role in recording research publications but also played a major role in the TARDis project and in the development of the University of Southampton Research Repository. It is now investigating the role of the library in the management and preservation of local data sources. Through the JISC funded Citation, Location, and Deposition in Discipline and Institutional Repositories (CLADDIER) project, the National Oceanographic Library is exploring the linking of its publications in the Institutional Repository with environmental data holdings [15]. The result will be a step on the road to a situation where active environmental scientists will to be able to move seamlessly from information discovery (location), through acquisition to deposition of new material, with all the digital objects correctly identified and cited. Experience at Southampton shows that a partnership between librarians and researchers is likely to give the best results – an experienced information manager/librarian is helpful in creating good citations for data entities (now given unique Digital Object Identifiers - DOIs) in the repository. Another example of the need for links between the Southampton Institutional Repository and a data archive is that of the eCrystals Crystal Structure

Report Archive [16]. Southampton is the home of the National Archive for Crystal Structures generated both by the Southampton Chemical Crystallography Group and by the Engineering and Physical Sciences Research Council (EPSRC) UK National Crystallography Service that is located on the Southampton campus. This raises questions as to which organization owns the long term responsibility for a national service which is created from a project by academics at the institution. In one model, this can be seen as just another strand of the institutional repository in capturing all intellectual assets – publications, data, learning objects and outputs such as proceedings and papers from workshops. The lessons learned from these examples will be valuable in establishing clear relationships and responsibilities between discipline based repositories and institutional repositories.

There are many other issues – such as those of provenance and preservation. In many research fields there are national repositories responsible for the curation and preservation of their scientific data. University libraries, on the other hand, may need to take responsibility for assisting with the curation and preservation of smaller scale data sets arising from the research of research groups or individual academics. The increasing importance of digital Institutional Repositories is giving an impetus to examine the associated preservation issues (Hitchcock *et al.*, 2005). Repository administrators will need to be supported in these underlying issues so that they can concentrate on their key goals of recording and providing access to scholarly output.

Conclusions

The advent of e-Science heralds a new and exciting world for the library world to be involved in. In both Europe and the USA there are now moves to develop a powerful infrastructure to support collaborative, multidisciplinary science. Such infrastructure is termed 'e-Infrastructure' in Europe and Cyberinfrastructure in the USA. One component of this infrastructure will be 'Grid' middleware that enables researchers to easily set up their own secure 'Virtual Organizations' linking research sites with whom they wish to share a variety of resources with controlled authenticated access. A second ingredient of this research infrastructure is of course the underlying research network that constitutes the academic research Internet. The last key ingredient of the research infrastructure is access to research results – both publications and data. Thus the e-Science revolution will put libraries and repositories centre stage in the development of the next generation research infrastructure.

Acknowledgements

The authors wish to thank Jim Gray, Stevan Harnad, Liz Lyon, Jean Paoli and Pauline Simpson for many informative discussions on the subject of this paper.

Notes

1. www.CombeChem.org
2. www.ukoln.ac.uk/projects/ebank-uk

3. www.SmartTea.org
4. www.soros.org/openaccess/
5. http://arxiv.org/
6. www.pubmedcentral.nih.gov
7. www.eprints.org/
8. The VERSIONS project: http://library-2.lse.ac.uk/versions/
9. The Wellcome Library: http://library.wellcome.ac.uk/node280.html
10. www.openarchives.org/OAI/openarchivesprotocol.html
11. http://archives.eprints.org/
12. www.opendoar.org/
13. http://ndg.nerc.ac.uk
14. www.slac.stanford.edu/spires/
15. http://claddier.badc.ac.uk
16. http://ecrystals.chem.soton.ac.uk/

References

Atkins, D. *et. al.* (Eds.) (2003), *Revolutionizing Science and Engineering Through Cyberinfrastructure*, Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure,, available at: www.nsf.gov/cise/sci/reports/atkins.pdf (accessed 4 April 2006).

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003), available at: www.zim.mpg.de/openaccess-berlin/berlindeclaration.html (accessed 4 April 2006).

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), "The semantic web", *Scientific American*, Vol. 284, No. 5, pp. 34-43.

Frey, J.G., De Roure, D. and Carr, L.A. (2002), "Publication at source: scientific communication from a publication web to a data grid", *Euroweb 2002 Conference: The Web and the Grid: From e-Science to e-Business*, Oxford, UK, December 17-18, 2002, available at: http://ewic.bcs.org/conferences/2002/euroweb/session3/paper3.pdf (accessed 4 April 2006).

Frey, J.G., Bradley, M., Essex, J.W., Hursthouse, M.B., Lewis, S.M., Luck, M.M., Moreau, L., De Roure, D., Surridge, M. and Welsh, A. (2003), "Combinatorial Chemistry and the Grid", in Berman, F., Fox, G. and Hey, T. (Eds.) *Grid Computing: Making the Global Infrastructure a Reality*, Wiley, Chichester, pp. 945-962.

Friedman, Th. (2005), *The World is Flat – A Brief History of the Twenty-First Century*, Farrar, Strauss and Giroux, New York, NY.

Harnad, S. and Hey, J. (1995), "Esoteric knowledge: the scholar and scholarly publishing on the Net", in: Dempsey, L., Law, D. and Mowat, I. (Eds.),

Networking and the Future of Libraries 2: Managing the Intellectual Record, Library Association Publishing, London, pp. 110-116.

Hey, J., Simpson, P. and Carr, L. (2005), "The TARDis Route Map to Open Access: developing an Institutional Repository Model", in Dobreva, M. and Engelen, J. (Eds.), *ELPUB2005 From Author to Reader: Challenges for the Digital Content Chain: Proceedings of the 9th ICCC International Conference on Electronic Publishing*, , Katholieke Universiteit Leuven, Leuven-Heverlee, Belgium, 8-10 June 2005, Peeters Publishing, Leuven, Belgium, pp. 179-182, available at: http://eprints.soton.ac.uk/16262/ (accessed 4 April 2006).

Hey, T. and Trefethen, A. (2003), "The data deluge", in Berman, F., Fox, G. and Hey, T. (Eds.), *Grid Computing: Making the Global Infrastructure a Reality*, Wiley, Chichester, pp. 809-824.

Hitchcock, S., Brody, T., Hey, J. and Carr, L. (2005), "Preservation for institutional repositories: practical and invisible", in *Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data (PV2005)*, The Royal Society, Edinburgh, Scotland, 21-23 November 2005, available at: http://eprints.soton.ac.uk/18774/ (accessed 4 April 2006).

Hughes, G., Mills, H., De Roure, D., Frey, J.G., Moreau, L., Schraefel, M.C., Smith, G. and Zaluska, E. (2004), "The semantic smart laboratory: a system for supporting the chemical e-Scientist", *Organic and Biomolecular Chemistry*, Vol. 2, No. 22, pp. 3284-3293.

Lynch, C. (2003), "Institutional repositories: essential infrastructure for scholarship in the digital age", *ARL BiMonthly Report*, No. 226, pp. 1-7; available at: www.arl.org/newsltr/226/ir.html (accessed 4 April 2006).

NIH News (3 February 2005), "NIH calls on scientists to speed public release of research publications", available at: www.nih.gov/news/pr/feb2005/od-03.htm (accessed 4 April 2006).

Schraefel, M.C., Hughes, G., Mills, H., Smith, G. and Frey, J.G. (2004a), "Making tea: iterative design through analogy"', in *Proceedings of the Conference on Designing Interactive Systems*, available at: http://eprints.soton.ac.uk/15879/01/schraefel-MakingTeaDIS04.pdf (accessed 4 April 2006).

Schraefel, M.C., Hughes, G., Mills, H., Smith, G., Payne, T. and Frey, J. (2004b), "Breaking the book: translating the chemistry lab book to a pervasive computing environment", in *Proceedings of the Conference on Human Factors (CHI)*, available at: http://eprints.soton.ac.uk/15878/01/schraefel_smartteaChi04.pdf (accessed 4 April 2006).

Segaller, S. (1998), *Nerds: A Brief History of the Internet*, TV Books, New York, NY.

Simpson, P. and Hey, J. (2005), "Institutional e-Print repositories for research visibility", in Drake, M. (Ed.), *Encyclopedia of Library and Information Science*, 2nd. ed., Dekker, New York, NY, available at: http://eprints.soton.ac.uk/9057/ (accessed 4 April 2006).

Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G. and Smith, M. (2003), “The DSpace Institutional Digital Repository System: Current Functionality”, in *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL’03)*, pp. 87-97.

Taylor, J.M. (2001), Talk given at UK e-Science Town Meeting, July, 2001.

Woodland, J. and Ng, J. (2006), "Too many systems, too little time: integrating an eprint repository into a University publications system”, in *VALA 2006 13th Biennial Conference and Exhibition*, Crown Towers Melbourne, Australia, 8 - 10 February 2006, Victorian Association for Library Automation, Melbourne, Australia, available at: http://espace.lis.curtin.edu.au/archive/00000618/ (accessed 4 April 2006).