36

The data deluge: an e-Science perspective

Tony Hey[1,2] and Anne Trefethen[1]

[1]*EPSRC, Swindon, United Kingdom,* [2]*University of Southampton, Southampton, United Kingdom*

36.1 INTRODUCTION

There are many issues that should be considered in examining the implications of the imminent flood of data that will be generated both by the present and by the next generation of global 'e-Science' experiments. The term *e-Science* is used to represent the increasingly global collaborations – of people and of shared resources – that will be needed to solve the new problems of science and engineering [1]. These e-Science problems range from the simulation of whole engineering or biological systems, to research in bioinformatics, proteomics and pharmacogenetics. In all these instances we will need to be able to pool resources and to access expertise distributed across the globe. The information technology (IT) infrastructure that will make such collaboration possible in a secure and transparent manner is referred to as the *Grid* [2]. Thus, in this chapter the term *Grid* is used as a shorthand for the middleware infrastructure that is currently being developed to support global e-Science collaborations. When mature, this Grid middleware

Grid Computing – Making the Global Infrastructure a Reality. Edited by F. Berman, A. Hey and G. Fox
 ISBN: 0-470-85319-0

will enable the sharing of computing resources, data resources and experimental facilities in a much more routine and secure fashion than is possible at present. Needless to say, present Grid middleware falls far short of these ambitious goals. Both e-Science and the Grid have fascinating sociological as well as technical aspects. We shall consider only technological issues in this chapter.

The two key technological drivers of the IT revolution are Moore's Law – the exponential increase in computing power and solid-state memory – and the dramatic increase in communication bandwidth made possible by optical fibre networks using optical amplifiers and wave division multiplexing. In a very real sense, the actual cost of any given amount of computation and/or sending a given amount of data is falling to zero. Needless to say, whilst this statement is true for any fixed amount of computation and for the transmission of any fixed amount of data, scientists are now attempting calculations requiring orders of magnitude more computing and communication than was possible only a few years ago. Moreover, in many currently planned and future experiments they are also planning to generate several orders of magnitude more data than has been collected in the whole of human history.

The highest performance supercomputing systems of today consist of several thousands of processors interconnected by a special-purpose, high-speed, low-latency network. On appropriate problems it is now possible to achieve sustained performance of several teraflop per second – a million million floating-point operations per second. In addition, there are experimental systems under construction aiming to reach petaflop per second speeds within the next few years [3, 4]. However, these very high-end systems are, and will remain, scarce resources located in relatively few sites. The vast majority of computational problems do not require such expensive, massively parallel processing but can be satisfied by the widespread deployment of cheap clusters of computers at university, department and research group level.

The situation for data is somewhat similar. There are a relatively small number of centres around the world that act as major repositories of a variety of scientific data. Bioinformatics, with its development of gene and protein archives, is an obvious example. The Sanger Centre at Hinxton near Cambridge [5] currently hosts 20 terabytes of key genomic data and has a cumulative installed processing power (in clusters – not a single supercomputer) of around 1/2 teraflop s^{-1}. Sanger estimates that genome sequence data is increasing at a rate of four times each year and that the associated computer power required to analyse this data will 'only' increase at a rate of two times per year – still significantly faster than Moore's Law. A different data/computing paradigm is apparent for the particle physics and astronomy communities. In the next decade we will see new experimental facilities coming on-line, which will generate data sets ranging in size from hundreds of terabytes to tens of petabytes per year. Such enormous volumes of data exceed the largest commercial databases currently available by one or two orders of magnitude [6]. Particle physicists are energetically assisting in building Grid middleware that will not only allow them to distribute this data amongst the 100 or so sites and the 1000 or so physicists collaborating in each experiment but will also allow them to perform sophisticated distributed analysis, computation and visualization on all or subsets of the data [7–11]. Particle physicists envisage a data/computing model with a hierarchy of data centres with associated computing resources distributed around the global collaboration.

The plan of this chapter is as follows: The next section surveys the sources and magnitudes of the data deluge that will be imminently upon us. This survey is not intended to be exhaustive but rather to give numbers that will illustrate the likely volumes of scientific data that will be generated by scientists of all descriptions in the coming decade. Section 36.3 discusses issues connected with the annotation of this data with metadata as well as the process of moving from data to information and knowledge. The need for metadata that adequately annotates distributed collections of scientific data has been emphasized by the Data Intensive Computing Environment (DICE) Group at the San Diego Supercomputer Center [12]. Their Storage Resource Broker (SRB) data management middleware addresses many of the issues raised here. The next section on Data Grids and Digital Libraries argues the case for scientific data digital libraries alongside conventional literature digital libraries and archives. We also include a brief description of some currently funded UK e-Science experiments that are addressing some of the related technology issues. In the next section we survey self-archiving initiatives for scholarly publications and look at a likely future role for university libraries in providing permanent repositories of the research output of their university. Finally, in Section 36.6 we discuss the need for 'curation' of this wealth of expensively obtained scientific data. Such digital preservation requires the preservation not only of the data but also of the programs that are required to manipulate and visualize it. Our concluding remarks stress the urgent need for Grid middleware to be focused more on data than on computation.

36.2 THE IMMINENT SCIENTIFIC DATA DELUGE

36.2.1 Introduction

There are many examples that illustrate the spectacular growth forecast for scientific data generation. As an exemplar in the field of engineering, consider the problem of health monitoring of industrial equipment. The UK e-Science programme has funded the DAME project [13] – a consortium analysing sensor data generated by Rolls Royce aero-engines. It is estimated that there are many thousands of Rolls Royce engines currently in service. Each trans-Atlantic flight made by each engine, for example, generates about a gigabyte of data per engine – from pressure, temperature and vibration sensors. The goal of the project is to transmit a small subset of this primary data for analysis and comparison with engine data stored in three data centres around the world. By identifying the early onset of problems, Rolls Royce hopes to be able to lengthen the period between scheduled maintenance periods thus increasing profitability. The engine sensors will generate many petabytes of data per year and decisions need to be taken in real time as to how much data to analyse, how much to transmit for further analysis and how much to archive. Similar (or larger) data volumes will be generated by other high-throughput sensor experiments in fields such as environmental and Earth observation, and of course human healthcare monitoring.

A second example from the field of bioinformatics will serve to underline the point [14]. It is estimated that human genome DNA contains around 3.2 Gbases that translates to only about a gigabyte of information. However, when we add to this gene sequence data, data on the 100 000 or so translated proteins and the 32 000 000 amino acids, the relevant

data volume expands to the order of 200 GB. If, in addition, we include X-ray structure measurements of these proteins, the data volume required expands dramatically to several petabytes, assuming only one structure per protein. This volume expands yet again when we include data about the possible drug targets for each protein – to possibly as many as 1000 data sets per protein. There is still another dimension of data required when genetic variations of the human genome are explored. To illustrate this bioinformatic data problem in another way, let us look at just one of the technologies involved in generating such data. Consider the production of X-ray data by the present generation of electron synchrotron accelerators. At 3 s per image and 1200 images per hour, each experimental station generates about 1 terabyte of X-ray data per day. At the next-generation 'DIAMOND' synchrotron currently under construction [15], the planned 'day 1' beamlines will generate many petabytes of data per year, most of which will need to shipped, analysed and curated.

From these examples it is evident that e-Science data generated from sensors, satellites, high-performance computer simulations, high-throughput devices, scientific images and so on will soon dwarf all of the scientific data collected in the whole history of scientific exploration. Until very recently, commercial databases have been the largest data collections stored electronically for archiving and analysis. Such commercial data are usually stored in Relational Database Management Systems (RDBMS) such as Oracle, DB2 or SQLServer. As of today, the largest commercial databases range from 10 s of terabytes up to 100 terabytes. In the coming years, we expect that this situation will change dramatically in that the volume of data in scientific data archives will vastly exceed that of commercial systems. Inevitably this watershed will bring with it both challenges and opportunities. It is for this reason that we believe that the data access, integration and federation capabilities of the next generation of Grid middleware will play a key role for both e-Science and e-Business.

36.2.2 Normalization

To provide some sort of normalization for the large numbers of bytes of data we will be discussing, the following rough correspondences [16] provide a useful guide:

A large novel	1 Mbyte
The Bible	5 Mbytes
A Mozart symphony (compressed)	10 Mbytes
OED on CD	500 Mbytes
Digital movie (compressed)	10 Gbytes
Annual production of refereed journal literature (~20 k journals; ~2 M articles)	1 Tbyte
Library of Congress	20 Tbytes
The Internet Archive (10 B pages) (From 1996 to 2002) [17]	100 Tbytes
Annual production of information (print, film, optical & magnetic media) [18]	1500 Pbytes

Note that it is estimated that printed information constitutes only 0.003% of the total stored information content [18].

36.2.3 Astronomy

The largest astronomy database at present is around 10 terabytes. However, new telescopes soon to come on-line will radically change this picture. We list three types of new 'e-Astronomy' experiments now under way:

1. *Virtual observatories*: e-Science experiments to create 'virtual observatories' containing astronomical data at many different wavelengths are now being funded in the United States (NVO [19]), in Europe (AVO [20]) and in the United Kingdom (AstroGrid [21]). It is estimated that the NVO project alone will store 500 terabytes per year from 2004.
2. *Laser Interferometer Gravitational Observatory (LIGO)*: LIGO is a gravitational wave observatory and it is estimated that it will generate 250 terabytes per year beginning in 2002 [22].
3. *VISTA*: The VISTA visible and infrared survey telescope will be operational from 2004. This will generate 250 GB of raw data per night and around 10 terabytes of stored data per year [23]. By 2014, there will be several petabytes of data in the VISTA archive.

36.2.4 Bioinformatics

There are many rapidly growing databases in the field of bioinformatics [5, 24]:

1. *Protein Data Bank (PDB)*: This is a database of 3D protein structures. At present there are around 20 000 entries and around 2000 new structures are being added every 12 months. The total database is quite small, of the order of gigabytes.
2. *SWISS-PROT*: This is a protein sequence database currently containing around 100 000 different sequences with knowledge abstracted from around 100 000 different scientific articles. The present size is of the order of tens of gigabytes with an 18% increase over the last 8 months.
3. *TrEMBL*: This is a computer-annotated supplement to SWISS-PROT. It was created to overcome the time lag between submission and appearance in the manually curated SWISS-PROT database. The entries in TrEMBL will eventually move to SWISS-PROT. The current release has over 600 000 entries and is updated weekly. The size is of the order of hundreds of gigabytes.
4. *MEDLINE*: This is a database of medical and life sciences literature (Author, Title, Abstract, Keywords, Classification). It is produced by the National Library of Medicine in the United States and has 11.3 M entries. The size is of the order of hundreds of gigabytes.
5. *EMBLnucleotide sequence database*: The European Bioinformatics Institute (EBI) in the United Kingdom is one of the three primary sites for the deposition of nucleotide sequence data. It contains around 14 M entries of 15 B bases. A new entry is received

every 10 s and data at the 3 centres – in the United States, United Kingdom and Japan – is synchronized every 24 h. The European Molecular Biology Laboratory (EMBL) database has tripled in size in the last 11 months. About 50% of the data is for human DNA, 15% for mouse and the rest for a mixture of organisms. The total size of the database is of the order of terabytes.

6. *GeneExpression database*: This is extremely data-intensive as it involves image data produced from DNA chips and microarrays. In the next few years we are likely to see hundreds of experiments in thousands of laboratories worldwide. Data storage requirements are predicted to be in the range of petabytes per year.

These figures give an indication of the volume and the variety of data that is currently being created in the area of bioinformatics. The data in these cases, unlike in some other scientific disciplines, is a complex mix of numeric, textual and image data. Hence mechanisms for curation and access are necessarily complicated. In addition, new technologies are emerging that will dramatically accelerate this growth of data. Using such new technologies, it is estimated that the human genome could be sequenced in days rather than the years it actually took using older technologies [25].

36.2.5 Environmental science

The volume of data generated in environmental science is projected to increase dramatically over the next few years [26]. An example from the weather prediction community illustrates this point.

The European Centre for Medium Range Weather Forecasting (ECMWF) in Reading, United Kingdom, currently has 560 active users and handles 40 000 retrieval requests daily involving over 2 000 000 meteorological fields. About 4 000 000 new fields are added daily, amounting to about 0.5 terabytes of new data. Their cumulative data store now contains 3×10^9 meteorological fields and occupies about 330 terabytes. Until 1998, the increase in the volume of meteorological data was about 57% per year; since 1998, the increase has been 82% per year. This increase in data volumes parallels the increase in computing capability of ECMWF supercomputers.

This pattern is mirrored in the United States and elsewhere. Taking only one agency, NASA, we see predicted rises of data volumes of more than tenfold in the five-year period from 2000 to 2005. The Eros Data Center (EDC) predicts that their data holdings will rise from 74 terabytes in 2000 to over 3 petabytes by 2005. Similarly, the Goddard Space Flight Center (GSFC) predicts that its holdings will increase by around a factor of 10, from 154 terabytes in 2000 to about 1.5 petabytes by 2005. Interestingly, this increase in data volumes at EDC and GSFC is matched by a doubling of their corresponding budgets during this period and steady-state staffing levels of around 100 at each site It is estimated that NASA will be producing 15 petabytes of data by 2007. The NASA EOSDIS data holdings already total 1.4 petabytes.

In Europe, European Space Agency (ESA) satellites are currently generating around 100 GB of data per day. With the launch of Envisat and the forthcoming launches of the Meteosat Second Generation satellite and the new MetOp satellites, the daily data volume generated by ESA is likely to increase at an even faster rate than that of the NASA agencies.

36.2.6 Particle physics

The BaBar experiment has created what is currently the world's largest database: this is 350 terabytes of scientific data stored in an Objectivity database [27]. In the next few years these numbers will be greatly exceeded when the Large Hadron Collider (LHC) at CERN in Geneva begins to generate collision data in late 2006 or early 2007 [28]. The ATLAS and CMS experiments at the LHC each involve some 2000 physicists from around 200 institutions in Europe, North America and Asia. These experiments will need to store, access and process around 10 petabytes per year, which will require the use of some 200 teraflop s^{-1} of processing power. By 2015, particle physicists will be using exabytes of storage and petaflops per second of (non-Supercomputer) computation. At least initially, it is likely that most of this data will be stored in a distributed file system with the associated metadata stored in some sort of database.

36.2.7 Medicine and health

With the introduction of electronic patient records and improvements in medical imaging techniques, the quantity of medical and health information that will be stored in digital form will increase dramatically. The development of sensor and monitoring techniques will also add significantly to the volume of digital patient information. Some examples will illustrate the scale of the problem.

The company InSiteOne [29] is a US company engaged in the storage of medical images. It states that the annual total of radiological images for the US exceeds 420 million and is increasing by 12% per year. Each image will typically constitute many megabytes of digital data and is required to be archived for a minimum of five years.

In the United Kingdom, the e-Science programme is currently considering funding a project to create a digital mammographic archive [30]. Each mammogram has 100 Mbytes of data and must be stored along with appropriate metadata (see Section 36.3 for a discussion on metadata). There are currently about 3 M mammograms generated per year in the United Kingdom. In the United States, the comparable figure is 26 M mammograms per year, corresponding to many petabytes of data.

A critical issue for such medical images – and indeed digital health data as a whole – is that of data accuracy and integrity. This means that in many cases compression techniques that could significantly reduce the volume of the stored digital images may not be used. Another key issue for such medical data is security – since privacy and confidentiality of patient data is clearly pivotal to public confidence in such technologies.

36.2.8 Social sciences

In the United Kingdom, the total storage requirement for the social sciences has grown from around 400 GB in 1995 to more than a terabyte in 2001. Growth is predicted in the next decade but the total volume is not likely to exceed 10 terabytes by 2010 [31]. The ESRC Data Archive in Essex, the MIMAS service in Manchester [32] and the EDINA service in Edinburgh [33] have experience in archive management for social science. The MIMAS and EDINA services provide access to UK Census statistics, continuous

government surveys, macroeconomic time series data banks, digital map datasets, bibliographical databases and electronic journals. In addition, the Humanities Research Board and JISC organizations in the UK jointly fund the Arts and Humanities Data Service [34]. Some large historical databases are now being created. A similar picture emerges in other countries.

36.3 SCIENTIFIC METADATA, INFORMATION AND KNOWLEDGE

Metadata is data about data. We are all familiar with metadata in the form of catalogues, indices and directories. Librarians work with books that have a metadata 'schema' containing information such as Title, Author, Publisher and Date of Publication at the minimum. On the World Wide Web, most Web pages are coded in HTML. This 'HyperText Markup Language' (HTML) contains instructions as to the appearance of the page – size of headings and so on – as well as hyperlinks to other Web pages. Recently, the XML markup language has been agreed by the W3C standards body. XML allows Web pages and other documents to be tagged with computer-readable metadata. The XML tags give some information about the structure and the type of data contained in the document rather than just instructions as to presentation. For example, XML tags could be used to give an electronic version of the book schema given above.

More generally, information consists of semantic tags applied to data. Metadata consists of semantically tagged data that are used to describe data. Metadata can be organized in a schema and implemented as attributes in a database. Information within a digital data set can be annotated using a markup language. The semantically tagged data can then be extracted and a collection of metadata attributes assembled, organized by a schema and stored in a database. This could be a relational database or a native XML database such as Xindice [35]. Such native XML databases offer a potentially attractive alternative for storing XML-encoded scientific metadata.

The quality of the metadata describing the data is important. We can construct search engines to extract meaningful information from the metadata that is annotated in documents stored in electronic form. Clearly, the quality of the search engine so constructed will only be as good as the metadata that it references. There is now a movement to standardize other 'higher-level' markup languages, such as DAML + OIL [36] that would allow computers to extract more than the semantic tags and to be able to reason about the 'meaning or semantic relationships' contained in a document. This is the ambitious goal of Tim Berners-Lee's 'semantic Web' [37].

Although we have given a simple example of metadata in relation to textual information, metadata will also be vital for storing and preserving scientific data. Such scientific data metadata will not only contain information about the annotation of data by semantic tags but will also provide information about its provenance and its associated user access controls. These issues have been extensively explored by Reagan Moore, Arcot Rajasekar and Mike Wan in the DICE group at the San Diego Supercomputer Center [38]. Their SRB middleware [39] organizes distributed digital objects as logical 'collections' distinct from the particular form of physical storage or the particular storage representation. A

vital component of the SRB system is the metadata catalog (MCAT) that manages the attributes of the digital objects in a collection. Moore and his colleagues distinguish four types of metadata for collection attributes:

- Metadata for storage and access operations
- Provenance metadata based on the Dublin Core [40]
- Resource metadata specifying user access arrangements
- Discipline metadata defined by the particular user community.

In order for an e-Science project such as the Virtual Observatory to be successful, there is a need for the astronomy community to work together to define agreed XML schemas and other standards. At a recent meeting, members of the NVO, AVO and AstroGrid projects agreed to work together to create common naming conventions for the physical quantities stored in astronomy catalogues. The semantic tags will be used to define equivalent catalogue entries across the multiple collections within the astronomy community. The existence of such standards for metadata will be vital for the interoperability and federation of astronomical data held in different formats in file systems, databases or other archival systems. In order to construct 'intelligent' search engines, each separate community and discipline needs to come together to define generally accepted metadata standards for their community Data Grids. Since some disciplines already support a variety of existing different metadata standards, we need to develop tools that can search and reason across these different standards. For reasons such as these, just as the Web is attempting to move beyond information to knowledge, scientific communities will need to define relevant 'ontologies' – roughly speaking, relationships between the terms used in shared and well-defined vocabularies for their fields – that can allow the construction of genuine 'semantic Grids' [41, 42].

With the imminent data deluge, the issue of how we handle this vast outpouring of scientific data becomes of paramount importance. Up to now, we 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 we will no longer have this luxury. We therefore need to automate the discovery process – from data to information to knowledge – as far as possible. At the lowest level, this requires automation of data management with the storage and the organization of digital entities. At the next level we need to move towards automatic information management. This will require automatic annotation of scientific data with metadata that describes both interesting features of the data and of the storage and organization of the resulting information. Finally, we need to attempt to progress beyond structure information towards automated knowledge management of our scientific data. This will include the expression of relationships between information tags as well as information about the storage and the organization of such relationships.

In a small first step towards these ambitious goals, the UK GEODISE project [43] is attempting to construct a knowledge repository for engineering design problems. Besides traditional engineering design tools such as Computer Aided Design (CAD) systems, Computational Fluid Dynamics (CFD) and Finite Element Model (FEM) simulations on

high-performance clusters, multi-dimensional optimization methods and interactive visualization techniques, the project is working with engineers at Rolls Royce and BAESystems to capture knowledge learnt in previous product design cycles. The combination of traditional engineering design methodologies together with advanced knowledge technologies makes for an exciting e-Science research project that has the potential to deliver significant industrial benefits. Several other UK e-Science projects – the myGrid project [44] and the Comb-*e*-Chem project [45] – are also concerned with automating some of the steps along the road from data to information to knowledge.

36.4 DATA GRIDS AND DIGITAL LIBRARIES

The DICE group propose the following hierarchical classification of scientific data management systems [46]:

1. *Distributed data collection*: In this case the data is physically distributed but described by a single namespace.
2. *Data Grid*: This is the integration of multiple data collections each with a separate namespace.
3. *Federated digital library*: This is a distributed data collection or Data Grid with services for the manipulation, presentation and discovery of digital objects.
4. *Persistent archives*: These are digital libraries that curate the data and manage the problem of the evolution of storage technologies.

In this chapter we shall not need to be as precise in our terminology but this classification does illustrate some of the issues we wish to highlight. Certainly, in the future, we envisage that scientific data, whether generated by direct experimental observation or by *in silico* simulations on supercomputers or clusters, will be stored in a variety of 'Data Grids'. Such Data Grids will involve data repositories together with the necessary computational resources required for analysis, distributed around the global e-Science community. The scientific data – held in file stores, databases or archival systems – together with a metadata catalogue, probably held in an industry standard relational database, will become a new type of distributed and federated digital library. Up to now the digital library community has been primarily concerned with the storage of text, audio and video data. The scientific digital libraries that are being created by global, collaborative e-Science experiments will need the same sort of facilities as conventional digital libraries – a set of services for manipulation, management, discovery and presentation. In addition, these scientific digital libraries will require new types of tools for data transformation, visualization and data mining. We return to the problem of the long-term curation of such data and its ancillary data manipulation programs below.

The UK e-Science programme is funding a number of exciting e-Science pilot projects that will generate data for these new types of digital libraries. We have already described both the 'AstroGrid' Virtual Observatory project [21] and the GridPP project [10] that will be a part of a worldwide particle physics Grid that will manage the flood of data to be generated by the CERN LHC accelerator under construction in Geneva. In other areas of

science and engineering, besides the DAME [13] and e-Diamond [30] projects described above, there are three projects of particular interest for bioinformatics and drug discovery. These are the myGrid [44], the Comb-*e*-Chem [45] and the DiscoveryNet [47] projects. These projects emphasize data federation, integration and workflow and are concerned with the construction of middleware services that will automatically annotate the experimental data as it is produced. The new generation of hardware technology will generate data faster than humans can process it and it will be vital to develop software tools and middleware to support annotation and storage. A further project, RealityGrid [48], is concerned with supercomputer simulations of matter and emphasizes remote visualization and computational steering. Even in such a traditional High Performance Computing (HPC) project, however, the issue of annotating and storing the vast quantities of simulation data will be an important aspect of the project.

36.5 OPEN ARCHIVES AND SCHOLARLY PUBLISHING

In the United Kingdom, the Higher Education Funding Council, the organization that provides core funding for UK universities, is looking at the implications of the flood of e-Science data for libraries on a 10-year timescale. In such a 10-year time-frame, e-Science data will routinely be automatically annotated and stored in a digital library offering the 'usual' digital library services for management, searching and so on, plus some more specialized 'scientific data'–oriented services such as visualization, transformation, other types of search engines and so on. In addition, scientific research in many fields will require the linking of data, images and text so that there will be a convergence of scientific data archives and text archives. Scientific papers will also routinely have active links to such things as the original data, other papers and electronic theses. At the moment such links tend to be transitory and prone to breaking – perhaps the research group Web address '~tony' stops working when Tony leaves and so on. The Open Archive Initiative [49], which provides software and tools for self-archiving of their research papers by scientists, addresses this issue to some extent, but this is clearly a large issue with profound implications for the whole future of university libraries. On the matter of standards and interworking of scientific digital archives and conventional repositories of electronic textual resources, the recent move of Grid middleware towards Web services [50, 51] is likely to greatly facilitate the interoperability of these architectures.

Scholarly publishing will presumably eventually make a transition from the present situation – in which the publishers own the copyright and are therefore able to restrict the group of people who can read the paper – to a model in which publishers are funded not for the paper copy but for providing a refereeing service and a curated electronic journal archive with a permanent URL. The difference between this model (proposed by Stevan Harnad [52]) and Paul Ginsparg's 'Eprint' archive for physics papers [53] is that Ginsparg's model is central and discipline-based, whereas Harnad's is distributed and institution-based. Both models depend on publishers to implement the peer review for the papers. Peer review is essential in order to identify signal from noise in such public archives. In Harnad's model, researchers' institutions pay 'publishers' to organize

the peer reviewing of their research output and to certify the outcome with their journal name and its established quality standard. The institutions' research output, both pre-peer review 'preprints' and post-peer review 'postprints', are archived in distributed, interoperable institutional Eprint archives. The Open Archives Initiative is providing a metadata harvesting protocol that could enable this interoperability. Using open source archiving software partly sponsored by the Budapest Open Access Initiative of the Soros Foundation, a growing number of universities in the United States and elsewhere are setting up Eprint Archives to provide permanent open access to their research. In addition to archiving their own research output, users also want to be able to search these archives for related works of others. Using the metadata associated with the archived paper, the OAI Metadata Harvesting Protocol [54] provides one solution to the problem of constructing suitable search engines. Any search engine produced in this manner will only be as good as the metadata associated with the papers [55], so strengthening and extending the metadata tagging and standards is a task of very high priority.

It seems just a question of time before scholarly publishing makes the 'Harnad Switch' – the outcome that Harnad has for a decade been describing as both optimal and inevitable. Authors actually want to maximize the impact and uptake of their research findings by making them accessible to as many would-be users as possible, rather than having them restricted, as they were in the paper era, to the minority of wealthy research libraries that can afford the access tolls. The Web has changed publishing forever and such a transition is inevitable. A similar transformation is likely to affect university libraries. The logical role for a university library in 10 years will surely be to become the responsible organization that hosts and curates (digitally) all the research papers produced by the university. It will be the university library that is responsible for maintaining the digital archive so that the '~tony' link continues to work for posterity. The Caltech Library System Digital Collections project [56] and the MIT DSpace project with HP [57] are two interesting exemplars of such an approach. There is also the interesting issue of how much responsibility individual universities would undertake for hosting and curating the scientific data produced by their researchers. Presumably, some universities would act as repositories for the scientific data for a number of university e-Science 'collaboratories', as well as acting as mirror sites for other organizations in the collaboration. Of course, particular communities will support specialized data archives – such as those of the EBI [24] and some national research organizations – and no doubt there will be commercial archives as well. An important issue not considered here is the question of ownership of data. Since much of the research in universities is funded by public bodies, there is clearly room for debate as to the ownership – and the curation costs!

36.6 DIGITAL PRESERVATION AND DATA CURATION

Generating the data is one thing, preserving it in a form so that it can be used by scientists other than the creators is entirely another issue. This is the process of 'curation'. For example, the SWISS-PROT database is generally regarded as the 'gold standard' for protein structure information [58]. Curation is done by a team of 25 full-time curators split

between the Swiss Bioinformatics Institute and the EBI. This shows how expensive the curation process is and why it will be necessary to address this support issue – involving extreme levels of automated, semi-automated and manual annotation and data cleansing. In addition, preservation of the data will be a crucial aspect of the work of a data repository. A recent EU/US study [59] recommended the establishment of a 'Data Rescue Centre' that would be concerned with research into the longevity of electronic data archives. The report envisaged that such a centre would examine the issues concerned with the refreshment, replication, repackaging and transformation of data and become a centre of much-needed expertise in these technologies.

There are many technical challenges to be solved to ensure that the information generated today can survive long-term changes in storage media, devices and digital formats. An introduction to the issues surrounding this problem has been given by Rothenberg [60]. To illustrate these issues we shall briefly summarize a novel approach to long-term preservation recently suggested by Lorie [61]. Lorie distinguishes between the archiving of data files and the archiving of programs. The archiving of programs is necessary in order that their original behaviour with the original data set can be reproduced in the future. For example, it is likely that a significant percentage of the scientific digital data to be preserved will be generated directly via some program P. A simple example is a spreadsheet program. In order to make sense of the data in the future, we need to save the original program P that was used to create and manipulate the data along with the data itself. Of course, in one sense the program P is just a bit stream like the data it produces – but the important difference is that the machine and the operating system required to run P may no longer exist. Lorie discusses the pros and cons of two proposed solutions to this problem: 'conversion' – copying files and programs to each new system as new systems are introduced – and 'emulation' – saving the data and the program as a bit stream along with a detailed description of the original machine architecture and a textual description of what the original program P should do to the data. Lorie then proposes a third approach based on specifying the program P in terms of instructions for a 'Universal Virtual Computer' (UVC). When archiving data, the UVC would be used to archive the methods that are required to interpret the stored data stream. For archiving a program, the UVC would be used to specify the functioning of the original computer. It is not clear which of these three approaches will turn out to be most feasible or reliable. Needless to say, a solution to these problems is much more than just a technical challenge: all parts of the community from digital librarians and scientists to computer scientists and IT companies need to be involved.

36.7 CONCLUDING REMARKS

From the above discussion, it can be seen that the coming digital data deluge will have profound effects on much of the current scientific infrastructure. Data from a wide variety of new sources will need to be annotated with metadata, archived and curated so that both the data and the programs used to transform can be reproduced in the future. e-Scientists will want to search distributed sources of diverse types of data and co-schedule computation time on the nearest appropriate resource to analyse or visualize their results.

This vision of Grid middleware will require the present functionality of both SRB [39] and Globus [62] middleware systems and much more. The present move towards Grid Services and Open Grid Services Architecture represents a unique opportunity to exploit synergies with commercial IT suppliers and make such a Grid vision a reality.

ACKNOWLEDGEMENTS

The vision of the Grid described in this chapter – with its emphasis on the access and the integration of distributed data resources combined with that of remote access to distributed compute resources – owes much to discussions with many people. We would particularly like to acknowledge the contributions to our understanding of these issues from Jim Gray, Jeff Nick, Bill Johnstone, Reagan Moore, Paul Messina and the Globus team in the United States and Malcolm Atkinson, Stevan Harnad, Jessie Hey, Liz Lyon, Norman Paton and Paul Watson in the United Kingdom. We are also grateful to Malcolm Read of JISC for his ever innovative support, to Sir Brian Follet for his early insight into the implications of e-Science for libraries and for universities and to John Taylor for both his vision for e-Science and for obtaining funding for the UK e-Science programme.

The authors are also grateful to David Boyd, Reagan Moore and Stevan Harnad for some helpful detailed comments on an earlier version of this chapter.

REFERENCES

1. Taylor, J. M., http://www.e-science.clrc.ac.uk.
2. Foster, I. and Kesselman, C. (eds) (1999) *The Grid: Blueprint for a New Computing Infrastructure*. San Francisco, CA: Morgan Kaufmann Publishers.
3. Allen, F. *et al*. (2001) BlueGene: A vision for protein science using a petaflop computer. *IBM Systems Journal*, **40**(2), 310–327.
4. Sterling, T. (2002) The Gilgamesh MIND processor-in-memory architecture for petaflops-scale computing. *ISHPC Conference*, Kansai, Japan, May, 2002.
5. Sanger Institute, Hinxton, UK, http://www.sanger.ac.uk.
6. Gray, J. and Hey T. (2001) In search of petabyte databases. Talk at *2001 HPTS Workshop Asilomar*, 2001, www.research.microsoft/∼gray.
7. EU DataGrid Project, http://eu-datagrid.web.cern.ch.
8. NSF GriPhyN Project, http://www.griphyn.org.
9. DOE PPDataGrid Project, http://www.ppdg.net.
10. UK GridPP Project, http://www.gridpp.ac.uk.
11. NSF iVDGL Project, http://www.ivdgl.org.
12. Rajasekar, A., Wan, M. and Moore, R. (2002) MySRB & SRB – components of a data grid. *11th International Symposium on High Performance Distributed Computing*, Edinburgh, Scotland, 2002.
13. DAME Project, www.cs.york.ac.uk/DAME.
14. Stuart, D. (2002) Presentation at *NeSC Workshop*, Edinburgh, June, 2002.
15. DIAMOND Project, http://www.diamond.ac.uk.
16. Lesk, M. (1997) *Practical Digital Libraries*. San Francisco, CA: Morgan Kaufmann Publishers.
17. Internet Archive, http:// www.archive.org.
18. Lyman, P. and Varian, H. R. (2000) *How Much Information?* UC Berkeley School of Information Management & Systems Report, http://www.sims.berkeley.edu/how-much-info.

19. NVO, http://www.nvo.org.
20. AVO, htpp://www.eso.org/avo.
21. AstroGrid, http://www.astrogrid.ac.uk.
22. LIGO, http://www.ligo.caltech.edu.
23. VISTA, http://www.vista.ac.uk.
24. European Bioinformatics Institute, http://www.ebi.ac.uk.
25. Hassard, J. (2002); private communication.
26. Gurney, R. (2002); private communication.
27. BaBar Experiment, www.slac.stanford.edu/BFROOT/.
28. LHC Computing Project, http://lhcgrid.web.cern.ch/LHCgrid.
29. InSiteOne Digital Image Storing and Archive Service, http://www.Insiteone.com.
30. Proposed e-Diamond Project, http://e-science.ox.ac.uk/. See also Chapter 41.
31. Neathey, J. (2002); private communication.
32. MIMAS Service, http://www.mimas.ac.uk.
33. EDINA Service, http://edina.ac.uk.
34. Arts and Humanities Data Service, http://www.ahds.ac.uk.
35. Xindice Native XML Database, http://xml.apache.org/xindice.
36. DAML+OIL, http://www.daml.org/2001/03/daml+oil-index.html.
37. Berners-Lee, T., Fischetti, M. (1999) *Weaving the Web*. New York: Harper Collins.
38. Rajasekar, A. K. and Moore, R. W. (2001) Data and metadata collections for scientific applications. *European High Performance Computing Conference*, Amsterdam, Holland, 2001.
39. The Storage Resource Broker, http://www.npaci.edu/DICE/SRB.
40. Dublin Core Metadata Initiative, http://Dublin core.org.
41. DeRoure, D., Jennings, N. and Shadbolt, N. Towards a semantic grid. *Concurrency & Computation*; (to be published) and in this collection.
42. Moore, R. W. (2001) knowledge-based grids. *Proceeding of the 18th IEEE Symposium on Mass Storage Systems and Ninth Goddard Conference on Mass Storage Systems and Technologies*, San Diego, April, 2001.
43. The GEODISE Project, http://www.geodise.org/.
44. The myGrid Project, http://mygrid.man.ac.uk.
45. The Comb-e-Chem Project, http://www.combechem.org.
46. Moore, R. W. (2001) Digital Libraries, Data Grids and Persistent Archives. Presentation at *NARA*, December, 2001.
47. The DiscoveryNet Project, http://www.discovery-on-the.net.
48. The RealityGrid Project, http://www.realitygrid.org.
49. Lagoze, C. and Van De Sompel, H. (2001) The open archives initiative: building a low-barrier interoperability framework, *JCDL '01*. Roanoke, Virginia: ACM Press, pp. 54–62.
50. Foster, I., Kesselman, C. and Nick, J. Physiology of the grid. *Concurrency and Computation*; (to be published) and in this collection.
51. Hey, T. and Lyon, L. (2002) Shaping the future? grids, web services and digital libraries. *International JISC/CNI Conference*, Edinburgh, Scotland, June, 2002.
52. Harnad, S. and Hey, J. M. N. (1995) Esoteric knowledge: the scholar and scholarly publishing on the Net, in Dempsey, L., Law, D. and Mowlat, I. (eds) *Proceedings of an International Conference on Networking and the Future of Libraries: Managing the Intellectual Record*. Bath, 19–21 April, 1995 London: Library Association Publications (November 1995), pp. 110–16.
53. Ginsparg e-Print Archive, http://arxiv.org.
54. Lynch, C. (2001) *Metadata Harvesting and the Open Archives Initiative*, ARL Bimonthly Report 217:1–9, 2001.
55. Lui, Z., Maly, K., Zubair, M. and Nelson, M. (2001) Arc – An OAI service provider for cross-archive searching, *JCDL'01*, Roanoke, VA: ACM Press, pp. 65–66.
56. The Caltech Library Systems Digital Collections Project, http://library.caltech.edu/digital/.
57. The MIT Dspace Project, http://www.mit.edu/dspace/.
58. SWISS-PROT, http://www.ebi.ac.uk/swissprot.
59. EU/US Workshop on Large Scientific Databases, http://www.cacr.caltech.edu/euus.

60. Rothenberg, J. (1995) Ensuring the longevity of digital documents. *Scientific American*, **272**(1), 42–7.
61. Lorie, R. A. (2001) Long Term Preservation of Digital Information. *JCDL '01*, Roanoke, VA, June, 2001.
62. The Globus Project, http://www.globus.org.