erpaTraining

The Selection, Appraisal and Retention of Digital Scientific Data

ERPANET / CODATA Workshop

Biblioteca Nacional, Lisbon

FINAL REPORT

The Selection, Appraisal and Retention of Digital Scientific Data

ERPANET / CODATA Workshop

Biblioteca Nacional, Lisbon
December 15-17, 2003

CONTENTS

Executive Summary

Rapid advances in technology are impacting the way scientists work, allowing greater amounts of digital data to be produced in the majority of scientific disciplines. These technological advances are also changing the way scientists interact, creating opportunities for collaborations across disciplines, institutions, and countries. The ever-increasing data that are generated through these advances require active curation to ensure their longevity. The international EPRANET/CODATA seminar examined the current state of practice of the selection, appraisal and retention among diverse scientific communities and discussed how archival concepts can best be applied to the management and long-term preservation of digital data.

The seminar was held from the 15th until the 17th of December 2003 at the Biblioteca Nacional in Lisbon, and brought together more than sixty-five researchers, data managers, information specialists, archivists, and librarians from thirteen countries to discuss the issues involved in making critical decisions regarding the long-term preservation of the scientific record. One of the major aims for this seminar was to provide an international forum to exchange information about data archiving policies and practices across different scientific, institutional, and national contexts. The seminar proved to be extremely successful in enabling discussions between scientific and archival communities. The seminar also highlighted some conceptual hurdles to overcome before effective collaboration between the diverse communities can take place.

There is strong agreement that data are the basis of the scientific endeavour, and that scientific data often have more than one life as scientific ideas advance and new concepts emerge. The reuse of existing digital data also maximizes initial investments. Thus it is essential to properly preserve these data, ensuring that data are appraised; that the related metadata are preserved with the data; and that awareness is raised among funding agencies of the importance of digital preservation activity.

There are varying levels of archival practice in use among the scientific disciplines based on the various needs and practices of the different communities. The space sciences, for example, have a long history of using selection and appraisal guidelines and have been active in establishing metadata and software standards within their discipline. But in other disciplines, the efforts have been less coordinated and sustained. More effective collaboration between the various scientific communities and archivists will be necessary to harmonise efforts.

The costs of capturing and storing data can vary depending on discipline. For example, the space and earth observation sciences capture huge amounts of data daily (often measuring in the terabytes) whereas digital data are generated on a smaller scale within the social sciences and many laboratory sciences. The selection, appraisal and retention needs of the various disciplines will reflect this diversity.

The terminology of archiving posed some confusion among the participants. Standard archiving terms - such as appraisal or record - have different meanings among the participants depending on their background, discipline, or area of practice. Before any real progress can be made in tackling the larger issues of selection, appraisal, and retention, the more practical issues of semantics must be addressed.

Publishing models differ among the scientific disciplines and, especially with the advent of digital publishing, impact the way that some disciplines carry out research and disseminate their findings.

The need to establish a basic appraisal framework for digital scientific data is universally agreed. The specific criteria and appropriate timing, however, are less easily defined.

This seminar is an important first step in the journey towards openness and collaboration between scientific disciplines, archivists, and other information specialists in the area of data curation and preservation. The seminar illustrated areas where each can learn from the others in establishing common frameworks and guidelines that will enable the effective selection, appraisal and long-term retention of digital scientific data.

Introduction

This international seminar examined the current state of practice of the selection, appraisal and retention of scientific data among diverse scientific communities and discussed how archival concepts can best be applied to the management and preservation of their digital data.

Seminar Setting

ERPANET co-hosted this event with CODATA - the Committee on Data for Science and Technology[1] at the Biblioteca Nacional (BN)[2] in Lisbon[3]. An evening reception was held at the BN where participants and speakers were able to discuss the issues in a more informal setting.

More than sixty-five researchers, data and records managers, archivists and librarians participated from thirteen countries, mainly from the EU member states, but also from Bangladesh, the United States, Canada, Norway and Sweden. The variety of countries represented gave an insight into activities at the national level while the participation of several representatives from international organisations provided a more global view on activity. There was a rich mix of scientific, archival, and library perspectives represented, both in terms of the speakers and the participants. This led to stimulating discussions.

Aims and objectives

One of the major aims for this seminar was to provide an international forum to exchange information about data archiving policies and practices across different scientific, institutional, and national contexts. The objectives were to:

- Use disciplinary and interdisciplinary case studies to assess the commonalities and differences among disciplines and determine the extent to which common selection, appraisal and long-term retention principles and policies can be identified and applied, regardless of discipline, and those that are unique to a discipline or a category of data.

- Identify and discuss the key scientific, technical, management, and policy considerations for the successful implementation of appraisal, selection and retention principles and policies, with particular attention to issues of efficiency, effectiveness, and the broad range of potential benefits – economic and other – that can be achieved.

- Provide a networking opportunity for seminar participants to meet with other researchers, data managers, information specialists, archivists, and science policy experts across disciplinary and national boundaries.

The seminar commenced with an introduction to CODATA activities related to archiving scientific data. This was followed by an overview of archival experiences in

[1] http://www.codata.org. A brief overview of CODATA has been included as an appendix to this document.
[2] http://www.bn.pt.
[3] ERPANET and CODATA thank Maria Inês Cordeiro of the Calouste Gulbenkian Foundation, Lisbon for her assistance.

the selection, appraisal and retention of digital records. An example of deriving the maximum economic benefit from the long-term retention of scientific data was then presented. These three talks helped to set the context of current archival activity and to identify potential benefits that the long-term preservation of digital scientific data may offer.

The seminar moved on to examine digital archiving activity in a variety of scientific disciplines. Through a series of disciplinary and interdisciplinary case studies, activities in the physical, biological, space, social and earth sciences were identified. The case studies attempted to illustrate commonalities and differences among various scientific disciplines by offering a range of archiving, end-user, and national perspectives. In many cases, more than one of these perspectives was explored.

A moderated plenary session allowed participants and speakers to engage in discussions on the issues raised during the presentations. The dialogue was lively and produced very interesting ideas.

As the seminar progressed, it became clear that the concept of appraisal was neither universally understood nor agreed by all participants. On the last day, several speakers volunteered to discuss appraisal more thoroughly to help clarify the issue.

This report will follow the structure of the seminar and provide a brief synopsis of the presentations and discussions that took place.

Opening Remarks

The seminar opened with remarks from ERPANET Director Seamus Ross, Biblioteca National Deputy Director Fernanda Campos and Pedro Fernandes from the Instituto Gulbenkian de Ciencia.[4] Seamus Ross encouraged participants representing disparate scientific communities to communicate and share their experiences with regards to selection, appraisal and retention of digital scientific data. Fernanda Campos, from the Biblioteca Nacional (BN), recognised the need to clarify and communicate roles and responsibilities to ensure the effective management and preservation of digital data. Pedro Fernandes outlined a Gulbenkian training programme to produce hybrid professionals known as bio-informaticians. The bio-informatician combines scientific expertise with archiving and data management skills and will be much in demand as scientific disciplines generate larger amounts of digital data that require proactive and efficient curation.

The opening talks emphasised the need to improve inter-sectoral collaboration in the identification of common strategies and frameworks for the selection, appraisal and retention of digital scientific data.

Introduction to Themes

CODATA's Scientific Data Archiving Activities

Since the 1960s, CODATA has been active in promoting data management across geographic and disciplinary boundaries. William L. Anderson, of the CODATA Task Group on Preservation and Archiving of Scientific and Technical Data in Developing

[4] Instituto Gulbenkian de Ciencia, http://www.igc.gulbenkian.pt.

Countries,[5] presented a general overview of CODATA's archiving activities. He stressed that by enabling re-use and repurposing of digital scientific data, their full value may be realised through new analysis. He emphasised that effective management of digital data is vital for its re-use. Anderson identified four main issues that will affect the management of digital scientific data. These are:

- changing scientific practices (i.e., the way scientists work and publish)
- increasing amounts of digital data being generated
- collaborating to develop policies to address this new environment
- making the right decisions regarding technical options and standards

The concept of maximising the value of digital data is one that came up frequently during the seminar. Anderson believes that scientific data are often meaningless outside the disciplines and communities of scientists that generate and use them. Several others echoed this sentiment during the seminar. Therefore, the scientific and archival communities must work towards facilitating the preservation of contextual information. Anderson recommended the adoption of the OAIS model to facilitate data management and stressed the need for further investigation into the potential long-term costs associated with archiving digital data. He then touched upon the need for persistent digital object identifiers in the quest for long-term preservation and re-use[6]. As we cannot predict what future users will want the data for, it will be very difficult to find the perfect solution to ensure that data are easily accessible and retrievable in the long-term, but a shared approach will certainly be more effective and efficient. Collaboration and communication, especially with developing countries, will be vital in the scientific community's approach to preserving access to scientific data assets[7]. Promoting and facilitating these collaborations are objectives of CODATA'a scientific data and information archiving activities.

An Archivist's Perspective of Appraisal of Digital Records

Terry Eastwood, from the University of British Columbia, expanded on some of the archiving themes raised in Anderson's talk. With the ever-increasing complexity of digital objects, long-term preservation becomes more and more difficult to solve. Eastwood believes that the challenges surrounding digital preservation are basically the same regardless of discipline - namely technical obsolescence and media fragility. As these are threats facing all disciplines, a shared approach makes perfect sense.

Eastwood then described some of the work done by the International Research on Permanent Authentic Records in Electronic Systems (InterPARES)[8] to ensure the long-term accessibility of digitally recorded memory. As appraisal is effectively a judgement on further preservation activity, he stressed the need for appraisal to

[5] More information on CODATA's Archiving Task Group can be found at http://www.codata.org.

[6] This theme will be further explored in an upcoming ERPANET seminar to be held in Dublin in June 2004

[7] CODATA will take part in a conference to be held in Beijing in 2004 to address technical issues and policies in developing countries

[8] InterPARES Project http://www.interpares.org/.

occur as early as possible in the life cycle of the digital object. In order for this to occur, appraisal criteria and policies must be defined and implemented. As the ability to place a digital object into context is crucial for its long-term usability, context has a great impact on the object's long-term value. Eastwood believes that context could be a factor used to assess the value of a digital object during the appraisal stage.

Eastwood feels that auditing and validating the authenticity of the digital objects are essential for their long-term value. Users need to know who created the digital objects and why. This is especially true for scientific data as researchers rely on documentary sources to generate new understanding and create new knowledge'[9]. As authenticity directly impacts the value of digital objects, it must be guaranteed over their entire life cycle. Automation of the auditing and validation processes may be of great benefit but will rely on the development of suitable software. This means that the scientific community, archivists, researchers and data managers will need to work closely with information specialists to create viable solutions. Authenticity is something that concerned the majority of participants at the seminar and will require greater investigation to produce solid guidelines and strategies.

Eastwood concurred with Anderson that further exploration into the economic feasibility of long-term preservation is necessary. He sees the knowledge divide between large and small organisations as particularly worrying and argues that the lack of money and awareness in small institutions holds them back from tackling digital preservation. According to Eastwood, large organisations should lead the way in the development of strategies and tools that can be adopted and adapted by smaller institutions.

In conclusion, Eastwood emphasised that the nature of the digital data should be irrelevant. Whether scientific or cultural, the management and preservation of digital data depends on the application of core concepts such as appraisal and authenticity.

Deriving Maximum Economic and Social Benefits from Public Investments in Preservation of Scientific Data

Authenticity certifies that digital data are trustworthy. Trust is essential for encouraging the reuse of data. Peter Weiss of the U.S. National Oceanic and Atmospheric Administration's National Weather Service pointed to the potential economic and social benefits to be gained from enabling free access and reuse of government-funded data sets.[10] The collection of digital data sets is generally expensive. Reuse allows their maximum benefit to be realised. Unlike the cost recovery model generally preferred in the European Union, the open and unrestricted access model used by the United States is thought to generate greater income in the long run. Weiss believes that this leads to greater social benefit and used the results of a study conducted by PIRA[11] to illustrate this claim.

Weiss used the example of the data sets of the U.S. National Weather Service to demonstrate how free access to data and their reuse enables greater economic and

[9] InterPARES 2 Project, http://www.interpares.org/ip2/ip2_index.cfm.

[10] National Oceanic and Atmospheric Administration (NOAA) at http://www.noaa.org; Peter Weiss' paper "Borders in Cyberspace: Conflicting Public Sector Information Policies and their Economic Impacts" can be seen at http://www.weathe.gov/sp/Bordersreport2.pdf.

[11] PIRA International study on the *Commercial Exploitation of Europe's Public Sector Information* in 2000. Final Report for the European Commission, Directorate General for the Information Society (2000) ftp://ftp.cordis.lu/pub/econtent/docs/commercial_final_report.pdf.

social benefits. These data sets have potential impact on several commercial industries, especially weather risk management, which is a fast-growing industry in the United States. The Weather Risk Management Association commissioned a report[12] to determine whether open access has impacted the growth of the weather risk management sector in the United States compared with Europe where such resources are not freely available. The results revealed that the new sector was booming in the United States but was not as successful in Europe. Weiss argued that this was due to the fact that people are generally not prepared to pay for information. This leads to two scenarios: the information is stolen or people simply do without it. Either way, little revenue is generated from the fee-based model. Enabling open and unrestricted access to government-funded data provides an opportunity for commercial enterprises to be built around these resources. These generate new uses of the data, employment opportunities, and new classes of tax revenues. In addition, openly accessible data offer benefits to developing countries that might otherwise not be able to tap into these resources. Reducing the digital divide is a major goal for many of the seminar participants and recent developments regarding the dissemination of scientific research in Europe appear to support this view[13].

Anderson, Eastwood and Weiss recognised that digital data have value beyond that for which they were created but stressed that their maximum potential can only be realised through effective curation. The disciplinary and interdisciplinary case studies that followed illustrated the current state of data management within a variety of scientific disciplines.

[12] PricewaterhouseCoopers. The weather risk management industry: survey findings for November 1997 to March 2002. Prepared for the Weather Risk Management Association (2002). http://wrma.org.

[13] Declaration on Access to research Data from Public Funding adopted on 30 January 2004 in Paris. Final Communique of the Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004. http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html.

Disciplinary Case Studies

Three disciplines - physical sciences, astronomy, and biological sciences - were selected to illustrate the varying selection, appraisal and retention practices in the scientific community.

Physical Sciences

Jürgen Knobloch, from the European Organisation for Nuclear Research (CERN)[14], introduced both the archiving and end-user perspectives for the physical sciences, using particle physics as an example. CERN's mandate includes the permanent preservation of scientific knowledge and accordingly, archiving activities have been a vital part of CERN's work for some time. Like Anderson, he emphasised that digital scientific data depend on the perpetuation of context to ensure their long-term value and usability. Without it, the data are essentially meaningless. He admitted that while it is current practice and optimal to have input from the original scientific team involved in any experiment, this is not generally practical for the long-term. It is therefore essential that context be preserved through the application of quality metadata.

Since 1964, an international group, "Particle Data Group"[15], jointly led by LBL, Berkeley and CERN, Geneva appraises and compiles high-level data and results in field of particle physics. The group maintains a database of cross-sections and particle properties data that is published in even-numbered years as "Review of Particle Physics" and is also made available on the web.

CERN generates huge amounts of raw data daily. Most of these data are discarded at an early stage of the dataflow, but much are retained to enable analysis and to verify the scientific process. In addition to preserving data generated during CERN's research activities, digitised plans and information about complex equipment must be kept, in many cases long after the equipment has ceased to be used, as is the case with the Large Electron-Positron (LEP) Collider. The LEP are maintained on a mass storage system. However, the LEP analysis will continue to be run as a computer museum system trying to keep the software operational with the current operating system. To emphasise the need to maintain equipment and its related information, Knobloch referred to the former use of punch cards at CERN and how they are now completely unreadable. The upkeep of hardware and equipment as well as ensuring that appropriate security measures are in place must all be factored into the overall costs of the long-term preservation of digital data.

Knobloch cited several challenges facing the physical sciences, particularly the particle physics community. The size and distribution of the particle physics community can be problematic. Currently, more than 5000 scientists worldwide use data held at CERN. By creating tools that enable distributed access via the web[16], CERN hopes to make their data more easily accessed and analysed. The next generation of particle physics experiments at the Large Hadron Collider LHC at CERN currently under construction will collect twelve to fourteen petabytes of data annually (the equivalent of a twenty kilometre high stack of CDs). Securing an

[14] http://public.web.cern.ch/public/.

[15] http://pdg.lbl.gov/.

[16] The web was invented at CERN in the early 90s and is now used e.g. by the QUERO system developed by B. Knuteson, LBL for public access of data from High Energy Physics experiments.

adequate and reliable CPU and data access resources to analyse these data will be of the utmost importance. CERN is actively leading grid projects for distributed data analysis, which should help to address the computing power issue and enable global access. Knobloch stressed that the volume of data being generated is not likely to decrease, so further research into affordable storage and management will be crucial to the long-term preservation and usability of data.

Knobloch illustrated that the physical sciences, especially particle and high-energy physics, have a long history of using technology in the capture and dissemination of their research. While competition drives many experiments, the particle physics community is open to sharing data and working together to solve common challenges surrounding the preservation of digital data.

Astronomy

Like CERN, the ground-based and space astronomy generates a huge volume of digital data daily. Françoise Genova, of the Centre de Données astronomiques de Strasbourg[17], provided the archiving perspective for the space sciences case study focusing on the International Virtual Observatory (IVO)[18] in astronomy. She began with a brief history of the Centre de Données astronomiques de Strasbourg, which was created over thirty years ago to deal specifically with electronic data generated from astronomical observation and to create added-value services. As astronomy is based on the long-term analysis of celestial phenomena, the records must be preserved indefinitely. By maintaining the data over the long-term, the maximum return on the initial investment can be realised. Maximising the value of digital data was mentioned in the opening session and was reiterated throughout the seminar.

As there are few commercial restraints on the astronomical sciences, data are not seen as commodities to be protected, but rather as resources to be shared. Accordingly, the astronomy community has focused on metadata and interoperability as a means of enabling queries across data sets to facilitate re-use. The adherence to open standards and interoperability has also led to observational data being linked with published results, which in turn enables new research. The federated approach to the description of data allows a homogeneous view to be obtained from heterogeneous data sets. The open environment fostered within the astronomy has had a knock-on effect spurring the publishers of astronomical research to make results quickly and easily accessible. Observational data are generally made available after a one-year proprietary period while results published in journals are available in full after three years.

The International Virtual Observatory can be described as “an enabling and coordinating entity to foster the development of tools, protocols, and collaborations necessary to realize the full scientific potential of astronomical databases in the coming decade” (NVO White Paper, 2000). The astronomy community benefits from the existence of a coordinating body, the International Virtual Observatory Alliance (IVOA),[19] which views interoperability as a major goal and actively promotes the adoption of standards across astronomy Virtual Observatory projects. For instance, the IVOA has proposed an XML schema known as VOTable[20] to be used for astronomical tables in the IVO. Genova cited the importance of using de facto

[17] http://cdsads.u-strasbg.fr/.

[18] http://www.euro-vo.org/.

[19] http://www.ivoa.net/pub/info/index.html.

[20] http://www.ivoa.net/xml/VOTable/.

standards and stressed that the value of collaboration and communication with other organisations must not be underestimated.

Alex Szalay, from the Department of Physics and Astronomy at the John Hopkins University[21], offered the end-user perspective for the astronomy case study. Szalay explained that science has changed in recent decades. Formerly empirical, scientific research is increasingly model based. Observational, analytical, computational and data exploration models are commonly employed today. He revealed that a shift in the way that scientists publish research has also taken place. Scientists no longer collect, analyse and then publish research but rather publish the raw data and then analyse the data to produce results.

Szalay stated that, on average, the amount digital data being generated in astronomy are doubling each year. The volume of scientific data being generated in most disciplines is becoming too large to analyse without the aid of computers. Another key example of this is genome research. Large, distributed data sets that grow daily are the norm in the astronomical sciences. Up to five terabytes of data are produced each night at some centres. Accessing and retrieving information from such large data sets requires that astronomical data be federated and well-organised in databases. There is no real commercial value associated with astronomical data, as noted by Genova, so free and open access does not present any major problems. In fact, he believes that the well-documented, temporal and spatial nature of astronomical science data sets makes them an ideal testbed for further research into free access. This is where organisations like the IVOA, which currently enables access to 200 terabytes of data held in fourteen countries, can offer valuable practical experience.

Szalay identified a revolution in the cost of modern scientific research. Until recently, equipment purchases such as large telescopes took up the majority of the research budget. However, software has now surpassed the cost of equipment. As noted above, this is due to the vast amount of data that requires analysis. He warns that if we gather too much data we may lose the ability to analyse them. Moore's law states that computer processing power will double every eighteen months,[22] which allows for the generation of increasing amounts of data. However, as databases continue to grow in size, requests for data will take longer and longer to produce results. Szalay cited the difference in the length of time it takes to transfer gigabytes versus terabytes of data from a server to a researcher's computer and believes that we need to achieve a balance between getting the best possible results in a realistic timeframe. Depending on the individual users' needs, what are the best results that can be generated from a data server in two hours, two weeks or two months? Szalay believes that by moving the analysis to the data rather than the data to the individual for analysis, optimal use of data stored in large databases can be achieved.

Overall, astronomy relies heavily on the long-term retention of data to carry out new research. As such, astronomy has been a leader in the development and use of standards and interoperability. Astronomy appears to have a very proactive approach to the management and accessibility of their data.

The proactive approach described in this session led to great discussion on several major points including metadata, archiving, appraisal and publishing. While the use of metadata is championed in the space sciences, some participants feel that more

[21] http://www.pha.jhu.edu/people/faculty/szalay.html.
[22] http://www.bbc.co.uk/dna/h2g2/A270028.

quality metadata could be captured. Concerns were expressed on how metadata could be retrospectively applied to older, yet still valuable, data sets. As preservation is not currently viewed as an essential part of digitised scientific projects by funders, who would finance such activity? In addition, the fact that metadata and appraisal appear to mean different things to archivists, records managers, and librarians than they do to the scientific researchers may cause problems in the widespread application of archival standards.

Biological Sciences

Meredith Lane, of the Global Biodiversity Information Facility (GBIF)[23], presented the archiving perspective for the biological sciences. GBIF was established to enable the digital capture and dissemination of data related to natural history specimens.

The biological sciences cover a vast range of information. For example, genome research focuses on the cellular level while ecosystem research examines entire communities of organisms. In between these two poles lies biodiversity, which deals specifically with species and genus level research. GBIF is involved in this area of the biological sciences.

It is important to bear in mind that no catalogue of all living organisms exists, even on paper. This means that great variations in terminology can exist from scientist to scientist within the biological sciences. To reconcile the variations and inconsistencies in synonym and scientific names alone would take about ten years to complete. In an attempt to address this problem, GBIF is currently working on the development of an Electronic Catalogue of Names of Known Organisms.

The majority of GBIF data relates to primary observations – most often from researching specimens in natural history museums. As noted below by Schürer with regards to the social sciences, film and sound recordings are becoming more and more common in the biodiversity research arena. To date, digitisation of data has been used to enhance access and usability and has largely occurred in the genome and ecosystem sub-disciplines rather than in biodiversity. In fact, about ninety-five percent of genome research is currently available in a digitised format. Biodiversity, on the other hand, has just about five percent of species and genus research available in a digitised format. Therefore, a large amount of retroactive data capture must take place to enable access to this information in a digital format. Lane acknowledges that the migration of legacy data will be difficult.

Some strategies have been developed in the biological sciences to address standards. For example, the establishment of the U.S. National Center for Biotechnology Information's Genbank[24] has helped to create standards that all digitised genomic data being published must meet. Genbank development involved input from publishers as well as the research communities and is similar in some ways to the model used in the space sciences. Lane explained that the range and complexity of the biological sciences makes the retrieval of gathered data difficult. The species level research is in a unique position to bridge the terminological gap between other biological sciences. This is due to the fact that all species are cell-based organisms and all live in ecosystems. However, further collaboration between sub-disciplines within the biological sciences will be vital in addressing the larger terminology problems.

[23] http://www.gbif.org/.

[24] http://www.ncbi.nlm.nih.gov/Genbank/index.html.

A major concern for GBIF is to help redress the distribution of data to make it more easily available to developing countries. As such, all GBIF data are freely and openly available via the Internet. Greater equality in resource distribution among nations was mentioned earlier in the seminar both by William Anderson and Peter Weiss.

As described below by Gutmann, GBIF feels that data should remain with the creator and be made accessible via a centralised site. This allows data to be updated as often as necessary on the creator's database without the need to update data held in a central repository. Enabling data mining across distributed genome, ecosystems and species databases with a single query is a major goal for GBIF. The ability to search across distributed data sets within disciplines was an aim cited by many throughout the course of the seminar.

Lane indicated that selection is generally project driven in the biological sciences. For example, rice specimens will be digitised for a project looking at rice crops. The biological sciences differ from most other disciplines in that the physical specimen lies at the heart of all research. Basically, the primary data are the original specimen and all other recorded information is simply metadata. Primary data in a digitised format are very useful for access, but as long as the original specimen is preserved in museums and labs, experiments can be carried out repeatedly. Lane realises, however, that this is impossible for many other scientific disciplines, such as those focusing on unique observations, where events cannot be re-created. Consequently, she emphasised the importance of applying quality metadata to digital objects.

Individual scientists must consider what will happen to their data when they leave a project; will others be able to understand its meaning? Lane believes that a philosophical change in the way that researchers think about the long-term preservation of the digital data they create is necessary. Without such a change, no long-term guarantees for the maintenance of digital data can be given. GBIF recognises the importance of funding bodies acknowledging the value of preservation and re-use of data. On this note, Lane expressed her concerns regarding the funding of retrospective metadata application to legacy data.

Weber Amaral, from the International Plant Genetics Resources Institute (IPGRI),[25] presented the end-user perspective for the biological science case study, but pointed out that the IPGRI is a data producer as well. The IPGRI deals mainly with phenotypes (seeds) and conduct in situ and ex situ research to enrich crop knowledge in developing countries. Their work is driven for social benefit rather than for commercial development.

The IPGRI currently holds eleven gene banks that contain over 600,000 phenotypes. Much of these data are held in a variety of distributed locations. For example, corn data are held in Mexico where they are produced. Passport data allow for the comparison between this distributed information. An internal impact study identified that eighty-one percent of IPGRI database users come from developing countries. Increasing use among and benefit to developing countries were cited as goals by many of the participants in the seminar. The biological sciences are actively engaged in making these goals reality.

Amaral explained that a good flow of data exists, but, as Lane mentioned, problems involving taxonomy are frequent. Most often this occurs when records with the same identifying name exhibit different traits or have different names but exhibit the same

[25] http://www.ipgri.cgiar.org/.

traits. While terminology can present challenges, IPGRI is dedicated to ensuring that the quality and validation of data are maintained - from the field to the end-user. All participants in the seminar agreed that authenticity is a vital element for the long-term use of digital data.

The IPGRI System-wide Information Network for Genetic Resources (SINGER)[26] provides access to information on the collections of genetic resources held by the CGIAR Centres. IPGRI is currently investigating the expansion of SINGER to enable querying across databases containing animal, aquatic and forest genetic resources. Data mining across sub-discipline data sets is also a goal for the GBIF. Amaral emphasised that access to high quality data must be combined with an easy-to-use, yet robust, system to facilitate widespread use. Amaral stressed that improved communication between all stakeholders, from farmers to policy makers, will be necessary to develop such a system.

Work in the biological sciences is focused on making the distribution of biological data more equally accessible to both the developed world and developing countries. This approach differs somewhat from other disciplines and presents a different set of challenges. Taxonomies appear to be a major concern for the biological sciences and the need for an agreed set of terms, not only between sub-disciplines but also on an interdisciplinary level, became clearly evident by the end of the seminar.

Discussion

Differences in the needs and practices of various scientific disciplines will make it difficult, if not impossible, to define a ‘one size fits all’ approach to selecting, appraising, and retaining scientific data. For example, the biological sciences require the preservation of the physical specimen for any related data to have meaning. However, the space sciences rely on the recording of unique events, such as solar and lunar eclipses, as the event itself can never be recreated.

The concept of publication described in this session led to some interesting debate on the various models employed by different disciplines. The space sciences generally have one year to analyse data gathered before they are released while social scientists have only three months to get information into the public domain or ten percent of funding can be withheld. But, the high volume of data that can be gathered and potential competition in some other disciplines (like the earth observation sciences) can have an impact on the publishing time scale. The Organisation for Economic Co-operation and Development (OECD) Committee for Scientific and Technological Policy at Ministerial Level has recently called for publicly funded scientific research data to be made openly available[27], as described by Weiss, which will no doubt impact the publishing of scientific research data in the European Union.

The question of when appraisal should take place was the focus of debate during this session. For many scientific disciplines, appraisal occurs with the granting of funds.

[26] http://www.ipgri.cgiar.org/system/page.asp?frame=programmes/sgrp/homesinger.htm.

[27] Final Communiqué of the Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004.
Declaration on Access to Research Data from Public Funding adopted on 30 January 2004 in Paris
http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html.

This implies that if the project is worth funding, it is worth retaining. For others, appraisal does not happen until the data are formally acquisitioned into an archive.

These case studies indicated that there is going to be a blurring of the scientist's role with that of the archivist (in managing the volume of research data being produced) and the publisher (with regard to making the results accessible). How can archivists, scientists, publishers, and funding agencies best work together to maximise the long-term retention and accessibility of scientific data results? Training hybrid professionals, such as the bio-informatician described by Pedro Fernandes in the opening session, may be one viable option.

Interdisciplinary Case Studies

The social sciences and earth observation sciences were identified as the interdisciplinary case studies, and representatives from these communities illustrated archiving practices in Europe and the United States.

Social Sciences

As with the disciplinary case studies, the social sciences recognise the value associated with the application of metadata to their digital data. Myron Gutmann, of the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan[28], explained that the social sciences have long been associated with tabulating and sharing information and that the standard of metadata used in social science surveys has been well defined for quite some time.

Although digital data are not generated on the same scale as some other disciplines, the amount is increasing annually. The U.S. government is one of the largest producers of social science data, followed by universities and private researchers. Like Eastwood and several others, Gutmann stressed the need to appraise the data from the earliest possible point in their life cycle. As social science data are gathered directly from human subjects, confidentiality is of key concern and challenges regarding data protection make social science data difficult to manage. Any information made public that could identify the human subject can nullify the validity of the data.

An interesting approach to selection has been adopted within the social sciences. If a digital record does not have sufficient metadata attached to it to adequately explain its context and relevance, it is rejected for long-term preservation. Maintaining context was identified as a priority among all of the scientific disciplines represented at the seminar. The approach taken by the social science community could be an efficient way of appraising data, both at the point of creation and retrospectively.

Generally, social science research data remain on the server of the creator but metadata are deposited with dissemination bodies such as the ICPSR to facilitate resource discovery. Gutmann explained that a significant shift has occurred in recent years from improving the input of data to improving the output of data. Accordingly, tools and strategies to improve the retrieval of data need to be investigated and developed.

Gutmann revealed that the archiving of digital data within the social science community is sporadic. Most government-funded U.S. social science researchers deposit their research in the U.S. National Archives and Records Administration and university archives. But, it seems that researchers are often more interested in securing funding for the next research project than in preserving the data of completed studies. This is by no means unique to the social sciences community and indicates that incentives must be identified and promoted to encourage digital content creators to deposit their research with archives for future access and reuse[29].

[28] http://www.icpsr.umich.edu/index.html.
[29] The ICPSR have developed a list of benefits to depositing data in archives. See http://www.icpsr.umich.edu/access/deposit/benefits.html for additional information.

Kevin Schürer, of the U.K. Data Archive (UKDA)[30], portrayed some of the concepts discussed by Gutmann from a U.K. perspective. Like Gutmann, Schürer acknowledged that social science data are subject to access regulations that do not have as great an impact on research results in other disciplines. The implementation of Freedom of Information Act[31] in the United Kingdom in January 2005 will have serious implications for some publicly funded data sets. Data Protection[32] must also be taken into account to protect the identity of study subjects. Effective data management will be essential to ensure compliance with both U.K. freedom of information and data protection policies.

Schürer noted that the format of data being generated in many social science studies is changing - not only from paper based to digital, but also to include a range of formats including video, audio tapes and images. He explained that the majority of social science digitisation projects currently funded in the United Kingdom attempt to improve access to data but neglect preservation. He sees the fact that funders do not take the long-term value of scientific research data into account as a major challenge to be overcome. Many others endorsed this sentiment during the course of the seminar.

Schürer alluded to DSpace and EPrints[33] repository initiatives as potential methods of organisational or self-archiving social science research. However, this would not in any way guarantee that the data retained are of high research value or ensure that adequate levels of metadata are applied. This is because anyone can set up their own repository and populate it with digital documents of their own choosing. Quality control could be enforced through the creation of a central repository, which is something that UKOLN[34] are investigating with the ePrints UK project. Institutional repositories have not yet been proven to increase the amount of research results made freely available. This is largely due to a lack of awareness and incentive to deposit. Schürer mentioned the new JISC funded Digital Curation Centre[35], which will work towards fostering trust in digital repositories and promoting the benefits of preserving digital scientific data.

The UKDA currently ingests about 150 data sets each year. While there has been an increase in the number of people accessing the data sets, there is no real way of telling whether or not actual use has increased. Schürer stressed here that it is important to differ between access and use. Schürer then discussed appraisal strategies for the social sciences in the United Kingdom The establishment of the National Digital Archive of Datasets[36] to house survey-type data has been beneficial in establishing set appraisal criteria for this type of research. At the UKDA, data are appraised using their collection policy. The policy determines the specific nature of data to be acquired and identifies any gaps in the collections that require filling. Criteria such as geographic coverage and the application of quality metadata are currently used by the UKDA to assess the merit of digital data entering the archive. At the moment, the UKDA rejects up to fifty percent of the data they are offered based on their collection policy. Ensuring that repositories have the right to reject

[30] http://www.data-archive.ac.uk/.
[31] Freedom of Information (FoI), http://www.parliament.uk/parliamentary_publications_and_archives/foi_introduction.cfm.
[32] Data Protection Act http://www.hmso.gov.uk/acts/acts1998/19980029.htm.
[33] DSpace http://www.dspace.org/index.html ; EPrints http://www.eprints.org.
[34] United Kingdom Office of Library Networking, http://www.ukoln.ac.uk.
[35] National Digital Curation Centre http://www.jisc.ac.uk/index.cfm?name=dcc_news_040204.
[36] National Digital Archive of Datasets (NDAD), http://www.ndad.ulcc.ac.uk.

data sets that fall outside their scope of collecting can help to avoid acquiring data that may be too costly, both financially and in terms of staff resources, to maintain. The concept of data rejection introduced an interesting perspective to the seminar.

Selection and appraisal appear to be more clearly defined in the social sciences than in other disciplines. This may be in part related to the fact that data protection affects the social sciences in ways that do not affect other disciplines. Both Gutmann and Schürer recognise the fact that funding will play a huge role in determining the amount of data that can be aggressively preserved. However, both feel that this will probably amount to a very small percentage of the actual social science data generated.

The Earth Observation Sciences

John Faundeen provided the U.S. perspective on work being done at the Earth Resources Observation Systems (EROS) Data Center of the U.S. Geological Survey (USGS)[37]. The EROS Data Center is the world's largest civilian data centre and deals mainly with satellite images. On average, one to two terabytes are captured daily. To develop an efficient management strategy for their digital data, an internal, multidisciplinary committee of archivists, managers and scientists was consulted. A twenty-question checklist, developed by the U.S. National Archives and Records Administration, has been incorporated by the USGS to assist in the appraisal of digital data. The completeness of metadata and data continuity are vital for the future comparison of data. As such, these areas are highlighted in the internal appraisal checklist. The internal appraisal tool allowed collections to be marked against 100 points. This distribution of points is at the discretion of the data manager. Therefore, concepts such as scientific relevance, metadata completeness, spatial and temporal resolutions could be weighted according to each manager's own opinion. Faundeen recognises that objectivity can be a challenge with this system but feels that it is a step in the right direction. Budget requirements were not included as part of the weighting within this earlier, internal USGS tool.

Faundeen stressed the importance of recognising that data have a definite life-cycle. The life-cycle of data ranges from creation to active use and distribution, through to preservation and eventually to disposal. The concept of disposal is one that must be accounted for in any digital archive strategy. By clearly defining the life-cycle for all digital data, resources can be used more effectively to maintain only the data that warrants long-term preservation. Schürer emphasised this point as well during his presentation. A more formalized scientific records appraisal tool,[38] built upon the lessons learned from the earlier internal one, was developed to help ensure that the disposal of scientific collections occur in a transparent manner and asks some of the following questions:

- Are the data of use to other researchers or government?
- Are the data understandable?
- Is there any major consequence if these data are lost?

If the answer to any of these questions is no, then the data may not be worthy of preservation.

As is the case for all of the scientific disciplines, appraisal is vital for the long-term usability of earth observation data. The importance of appraisal was initially emphasised by Terry Eastwood and echoed by many others throughout the seminar. The development of an easy-to-use checklist to assist in the appraisal process proved to be an interesting concept to the majority of seminar participants.

Luigi Fusco, of the European Space Agency (ESA), discussed the European perspective for earth observations. Fusco described the work being carried out at the ESA European Space Research Institute[39]. Earth observation data are great in

[37] http://edc.usgs.gov/.
[38] http://edc2.usgs.gov/government/RAT/tool.asp.
[39] http://www.esa.int/export/esaSA/earth.html.

volume and range from the local to the global in context. Fusco described the Global Monitoring of Environmental and Security project[40], in which the 'ESA and the European Commission are already working closely together to make sure that their respective programmes will allow the research, development and operational user communities across Europe to unite in a coordinated effort to establish an autonomous European global monitoring capability for environment and security purposes by 2008'[41]. Currently, there is no mandate to preserve earth observation mission data at the European level and the responsibility falls under the remit of the individual mission owner. Coordinating efforts on standards and approaches to preserve the most valuable European earth observation data will be required to develop the solid infrastructure needed for the GMES project to enable timely access to data. Like Szalay, Fusco stressed that the computer should move to the data rather than the data to the computer for analysis.

Other collaborative initiatives underway at the ESA include the Charter on Space and Major Disasters (which is an agreement to share data on major disasters) and Envisat (an earth observation program that produces up to 400-500 terabytes of data per year). These projects generate large quantities of digital data that are potentially of great value to research scientists in the earth observation community and beyond. The ESA Oxygen initiative makes a large-scale attempt to blend and feed these data sets. Collaboration across communities has occurred and will hopefully produce high-level products. This experience makes the earth science user community a valuable partner in the creation of tools and applications.

Fusco stated that the grid could be very useful to the digital library community. The grid addresses the major digital library architecture requirements—openness, scalability, security and quality. He proposed a testbed be established combining the grid, earth observation data and the digital library architecture to test this theory further.

The earth sciences produce a very high volume of data. While this is also true of the space and physical sciences, it appears that the earth sciences are very active in inter-sectoral collaborative initiatives with regards to making their data accessible and manageable over the long-term. By working with researchers, archivists and data managers, the earth observation sciences have illustrated the potential value of interdisciplinary cooperation.

Discussion

This session urged participants to consider the lifespan of digital data. We cannot afford to preserve everything being created for perpetuity. Accordingly, appraisal and retention policies will be vital to ensure that funds are spent on the maintenance and preservation of data and related records, including metadata, that are of high value rather than spread too thinly over a massive volume of data that is of little consequence. The application of an appraisal tool could be beneficial in this process but should ideally include budget requirements. Participants were encouraged to consider that although data may no longer have value within one community, they could have potential value for other research communities. By scheduling the disposal of data and related records and communicating these plans with others, there is a finite period where other research communities can express their interest in taking over responsibility for the digital data. This might be done through listserv

[40] http://earth.esa.int/gmes/.

[41] *Ibid.*

alerts. The collaborative approaches used by the USGS and ESA offer models that could be adopted by other organisations in a variety of disciplines.

Other questions emerging from this session focussed on the concept of free and open access. Kevin Schürer argued that there is no such thing as free data. Data made free at the point of access still required financial input by those developing the service. Could such expenses be absorbed as part of operational costs? Much remains to be determined, but collaboration between the digital library and scientific communities is a very encouraging first step.

Plenary Session

Gail Hodge, of Information International Associates Inc, moderated the plenary session. Prior to opening the plenary session to the participants, Hodge summarized the major issues she saw emerging from the presentations and case studies.

Hodge emphasized that the digital world is different and that new challenges will require greater collaboration and sharing of experiences. She agreed with the general consensus that the application of quality metadata will be essential for any long-term use of scientific digital data, especially with regards to ensuring context.

Hodge concurred with Weiss and others that digital scientific data are of economic and social benefit, therefore the preservation of these data will be extremely important. She believes that it is generally more common to hear of preservation failures than successes. Organisations like ERPANET can help to dispel this belief by making success stories and best practice more widely available.

Like Anderson, she stated that future uses of archived data cannot be anticipated. Therefore, it will be difficult, if not impossible, to guarantee that we can meet all potential users' needs.

Finally, Hodge touched upon the differences in terminology that appear to exist between the various scientific disciplines and archivists. The presentations and ensuing discussions revealed that terms such as archive, metadata and record could have very different meanings across disciplines. Hodge then opened the floor to discussion.

One of the first issues raised reflected this disparity between archivists and scientists regarding the concept of archival terms. It was felt that this misunderstanding would affect the long-term preservation of the context of digital data. As such, the transfer of knowledge from archivists to scientists, including terminology, will be crucial.

Another issue that emerged in the plenary session was the fact that the dissemination of research results varies among scientific disciplines. For example, the social sciences are bound to share their results as a stipulation in funding contracts, otherwise up to ten percent of the funding can be withheld. However, this is not the case with the physical sciences.

This escalated to a discussion on publication models and how changes in the way that research results are disseminated will impact the way the scientists work. For instance, the health research community strives to keep results private until after publication whereas in the astronomical sciences privacy is less of an issue. Obviously the commercial value of data and competition come into play for some scientific disciplines and will require individual approaches. It was suggested that national legislation or policies might be necessary to regulate access to research results. For example, in Sweden scientists generally have the right to access raw and processed data but access can be restricted to protect economic interests. While this is generally an exception to the Principle of Public Access, restrictions under the 1980 Secrecy Act can impact access to data for up to seventy years.

Electronic publishing is changing the way that scientists work. Peer review is a pillar of the scientific process. In most cases, it is required prior to publication. And journal publications are an important factor in determining tenure for many academics. New electronic publishing initiatives, such as the Public Library of Science, are changing the way researchers publish and disseminate their research.

The concept of roles and responsibilities were further explored in the plenary session. Should scientists or archivists care for the long-term preservation of digital scientific data? Does the responsibility lie with the funding body that finances a research project or with the scientist or institution that carries out the research? What happens if no suitable archive currently exists to curate and preserve the data? There were no real answers to these questions, but there was common agreement that cooperation between scientists and archivists in addition to the application of high-level standards from the outset of any scientific research project would be beneficial. This would at least ensure that the data could be easily accepted into an appropriate digital archive should one exist. The OAIS model was mentioned during this session as a potential framework that could be adopted by the scientific community to ameliorate the situation. The introduction of incentives and rewards for adopting sound archival practices was suggested.

The fact that funders do not see the preservation of digital data as a necessity was something that several of the participants and speakers touched upon during the presentations. Funding bodies rarely view data management or secondary research projects, based on the re-use of data, as valuable enough to warrant financing. Therefore, raising the awareness among the funding community of the value of reusing data was seen as a major area requiring action following the seminar.

The participants debated the costs associated with long-term storage of digital data. Some believe that the costs will become cheaper as disks become denser and less expensive to buy. However, the cost of tapes do not seem to be getting any cheaper, which will affect disciplines where digital data are created in very large quantities such as the space and earth observation sciences. The cost of human input to the processes of migration and data cleansing were also cited as areas that would require expenditure.

Participants were concerned about the creation of ‘deep archives’ where preservation resources are expended on objects that are never accessed. Discussions revealed that archivists might be more pragmatic than scientists when it comes to the selection, appraisal and disposal of digital data. As such, it will be necessary to achieve a balance between what we can feasibly retain and what can be realistically disposed of. The possibility of re-appraising digital data over their lifecycle was explored.

The question of when a digital object should enter the archive was another area of dispute. Some felt that the awarding of funding justified deposit with an archive for long-term preservation while others felt that this was not, in itself, sufficient reason. Use among social science data sets has revealed that some data are used again and again while other data are never accessed. Therefore, linking long-term value to initial funding could be problematic.

The question of objectivity was discussed. Many participants believe that raw data are objective while processed data are biased. However, determining which level of data to archive depends greatly on the scientific discipline. There was a suggestion that the lowest level of data plus any additional information needed to understand the data should be recorded, for example raw data plus calibrations for the earth observation sciences.

By the end of the plenary session, it became clear that the overall concepts of appraisal, selection and retention of digital scientific data are too complex to solve in a single seminar. Thus, the problems should be tackled in smaller, more manageable projects and events. Semantic, institutional, and disciplinary differences must be addressed before any further discussion can be effectively undertaken.

Panel Discussions on Appraisal

As the seminar progressed, it became clear that the concept of appraisal was neither universally understood nor agreed by all participants. On the last day, an interdisciplinary panel featuring Kevin Schürer, Terry Eastwood, Jürgen Knobloch and John Faundeen explored appraisal in more detail. Ultimately, the panel agreed that digital data are different from data stored on analogue media. They also agreed that the concept of long-term preservation should not be confused with short-term access.

The development of an acquisition policy in all digital archives was recommended as a way of ensuring that data accepted into an archive comply with the organisation's overall mission. Acquisition policies enhance the transparency of the appraisal process and provide solid reasons for the rejection of data into the archive as well as the eventual disposal of data. The panel suggested that an acquisition committee be established for each digital archive to provide a greater level of objectivity and accountability.

The general consensus of the panel was that the appraisal process should be collaborative, involving both scientists and archivists and that appraisal should take place as early in the life cycle of digital data as possible.

OAIS Overview

The benefits of the OAIS model were referred to several times throughout the seminar. Donald Sawyer of the National Space Science Data Center[42] presented a brief overview of the model. The fact that the model is generic allows it to be easily applied across disciplines. It could also be of great assistance in overcoming terminological differences that exist among scientific disciplines as it offers a common language in which to communicate needs. Sawyer stated that appraisal might best be associated with the ingest area of the OAIS model[43]. Others felt that appraisal stemmed more naturally from the management and administration area. Others still felt that appraisal occurs outside the model altogether.

As a high-level conceptual model, the OAIS suggests no concrete actions to be followed. Therefore, organisations must carefully consider their own individual needs and requirements using the model as a point of reference. The widespread, international acceptance of the OAIS is a convincing argument for the adoption of the model within the scientific community. For a more detailed overview of the OAIS model and its potential benefits, please refer to the documentation produced following the ERPANET OAIS training seminar held in Copenhagen in 2002[44].

[42] http://nssdc.gsfc.nasa.gov/.

[43] Consultative Committee for Space Data Systems (January 2002), *CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS).* Blue Book. Issue 1, 1-1. http://www.ccsds.org/documents/650x0b1.pdf.

[44] www.erpanet.org→Products and services→erpaTraining→OAIS Training Seminar

Conclusions

A main goal of this seminar was to identify any commonalities and differences that exist with regards to the selection, appraisal and retention of digital scientific data between various scientific disciplines[45].

Several commonalities were highlighted during the seminar. These, for the most part, involved maximising the value of digital resources. It was generally agreed that this was best achieved through the rigorous reuse of digital data sets. It was universally agreed that context would be of crucial importance in enabling reuse of digital data and that this could only be guaranteed through the application of quality metadata. As this will involve greater financial investment in the initial creation or retrospective documentation of digital data sets, all felt that it would be necessary to increase awareness of the value of the reuse of data among funding bodies. Many participants felt that the lobbying of funding bodies to include data curation costs in the financing of any scientific project producing digital data should be actively pursued as a result of the seminar.

The seminar illustrated that there is a great range of archival activity being undertaken in the various scientific disciplines. While different disciplines focus their activity in different areas of the archival life cycle, all have valuable experience to share. Through improved communication and collaboration, a great deal can be learned from each other. Astronomy showed that metadata and interoperability can have a major impact on the accessibility of data. They also demonstrated the value of disciplinary cooperation in the creation and adoption of standards and strategies. The social sciences suggested that metadata can be used as means of appraising the long-term value of digital data. In addition, they provided valuable insights into the concepts of rejecting or disposing of digital data. The physical sciences demonstrated the importance of retaining contextual information with digital data for their long-term value and reuse. The biological sciences also outlined the value of maintaining context and demonstrated the potential social benefits of making digital scientific data more widely accessible, especially among developing countries. The earth observation sciences showed that interdisciplinary collaboration can be of great benefit as was seen with the development of the USGS Scientific Records Appraisal Tool[46]. The archival community demonstrated that appraisal is effectively a judgement on the value of digital data. As such, appraisal can have a huge impact on justifying future preservation activity being carried out on the digital data. The potential benefits of the widespread adoption of the OAIS model were also illustrated by the archival discipline.

Areas that proved to be quite different among the various scientific disciplines included publishing models and dissemination strategies, the commercial value of data and the volume of data produced. The seminar concluded that further investigation into long-term storage and preservation costs will be of benefit to all scientific disciplines, regardless of the amount of data held. In addition, a

[45] Various issues in the long-term retention of scientific and technical data were identified and addressed by the U.S. National Research Council (NRC); see NRC. 1995. *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources*, National Academy Press (NAP), Washington, D.C., available at http://books.nap.edu/catalog/4871.html, and NRC. 1995. *Study on the Long-Term Retention of Selected Scientific and Technical Records of the Federal Government: Working Papers*, NAP, Washington, DC.

[46] http://edc2.usgs.gov/government/RAT/tool.asp.

collaborative effort will be required to produce some kind of common vocabulary or definition of terms. Appended to this document, readers will find a list of definitions prepared by CODATA and the National Research Council of the U.S. National Academies to help start this process. Overcoming the terminological difficulties will be an important step in opening the channels of interdisciplinary communication and collaboration.

This seminar has proved to be an essential first step in highlighting areas where disciplines can learn from each other in the development of solutions regarding the selection, appraisal and retention of digital scientific data.

Appendix One: Seminar Programme

Monday, 15 December

14:00 Opening Remarks by Seminar Chairs and Local Sponsors
Seamus Ross, Director, HATII (University of Glasgow) & ERPANET,
William Anderson, Chair, CODATA Data Preservation Task Group
Fernanda Campos, Director, Biblioteca National and Professor Pedro Fernades, Instituto Gulbenkian de Ciencia

14:15 Introduction to CODATA Archiving Activities
William Anderson, Chair, CODATA Data Preservation Task Group

14:30 Keynote 1: What Archivists Have Learned about Appraisal of Digital Records
Terry Eastwood, University of British Columbia, Canada

15:00 Keynote 2: Deriving the Maximum Potential Scientific, Economic and Social Benefits from Public Investments in the Long-term Retention of Scientific Data
Peter Weiss, National Weather Service, USA

15:30 BREAK

15:45 Disciplinary Case Study 1: Physical Sciences
- Archiving perspective: *Jürgen Knobloch, CERN, Switzerland*

16:30 Disciplinary Case Study 2: Biological Sciences
- Archiving perspective: *Meredith Lane, GBIF Secretariat, Denmark*
- High-value user perspective: *Weber Amaral, International Plant Genetics Research Institute, Rome*

17:15 Disciplinary Case Study 3: Space Sciences
- Archiving perspective: *Francoise Genova, Strasbourg Astronomical Data Centre, France*
- High-value user perspective: *Alex Szalay, The Johns Hopkins University, USA*

18:00 General Discussion of Disciplinary Case Studies

18:30 Adjourn

18:45 Reception

Tuesday, 16 December

09:00 Interdisciplinary Case Study 1: Social Sciences

- U.S. Archiving perspective: *Myron Gutmann, ICPSR, USA*
- European Archiving perspective: *Kevin Schürer, UK Data Archive*

10:00 Interdisciplinary Case Study 2: Earth & Environmental Sciences

- U.S. Land Remote Sensing Archive: *John Faundeen, USGS Eros Data Center, USA*
- Earth Observation Archives in Virtual Digital Libraries and GRID Infrastructures: *Luigi Fusco, European Space Agency, Italy*

11:00 Break

11:30 General Discussion of Interdisciplinary Case Studies

14:00 Plenary Discussion
Moderator: *Gail Hodge, Information International Associates, USA*

- Common/unique long-term appraisal and selection guidelines and long-term retention policies
- Scientific, technical, management, and policy considerations for successful implementation

Wednesday, 17 December

9:00 Wrap-up Plenary Discussion
Chairs: *Seamus Ross, Director, HATII (University of Glasgow) & ERPANET, William Anderson, Chair, CODATA Data Preservation Task Group*

- Overview of appraisal John Faundeen, Terry Eastwood, Jürgen Knobloch, Kevin Schürer
- Discussion

11:30 Overview of OAIS model, Donald Sawyer, National Space Sciences Data Center, USA

12:15 Closing Remarks from Seminar Co-Chairs

12:30 End of Meeting

Appendix Two: List of Participants

Weber Amaral	IPGRI, Italy
William Anderson	CODATA Data Preservation Task Group, USA
Kathleen Arntz	National Archives & Records Administration, USA
Renata Arovelius	Swedish University of Agricultural Sciences (SLU), Sweden
Bernard Avril	University of Bergen – SMR, Norway
Rosa Bela Azevedo	IAN/TT, Portugal
Jose Borbinha	Biblioteca Nacional, Portugal
Georg Buechler	ERPANET, Switzerland
Fernanda Campos	Biblioteca Nacional, Portugal
Kathleen Cass	CODATA, France
Virgilio Castillo	United Nations
Maria José Chaves	IAN/TT, Portugal
Ian Cording	Pfizer Ltd, UK
David Corney	CCLRC Rutherford Appleton Lab, UK
Rosário Costa	Observatório da Ciência e do Ensino Superior, Portugal
Melissa Cragin	University of Illinois at Urbana-Champaign, USA
Jeffrey Darlington	The National Archives, UK
Joy Davidson	ERPANET, UK
Michael Day	UKOLN, UK
Acácio Lopes de Sousa	IAN/TT, Portugal
Janet Dodson	Glaxo Smith Kline, USA
Peter Dukes	Medical Research Council, UK
Terry Eastwood	The University of British Columbia, Canada
Julie Esanu	CODATA/The National Academies, USA
John Faundeen	U.S. Geological Survey, EROS Data Center, USA
Pedro Fernandes	Gulbenkian Institute, Portugal
Maria Isabel Goulo Ferreira	IAN/TT, Portugal
Maria José Fidalgo	Biblioteca Nacional, Portugal
Luigi Fusco	ESA / ESRIN, Italy
Helena Galhardas	INESC / IST, Portugal
Rosa Galvão	Biblioteca Nacional, Portugal
Kari Lien Garnes	Bergen University Library, Norway
Françoise Genova	CDS, Observatoire de Strasbourg, France
Jens Goessner	Learninglab Lower Saxony, Germany

Zélia Gomes	IAN/TT, Portugal
Elizabeth Griffin	Dominion Astrophysical Observatory, Canada
Myron Gutmann	The University of Michigan, USA
Peter Harper	National Cataloguing Unit for the Archives of Contemporary Scientists, UK
Cecília Henriques	National Archives of Portugal, Portugal
Gail Hodge	Information International, USA
Delphine Jensen	European Investment Bank, Louxembourg
Jürgen Knobloch	CERN, Switzerland
Meredith Lane	Global Biodiversity Information Facility, Denmark
Paulo Jorge Leitão	Biblioteca Municipal de Almada, Portugal
W. Christopher Lenhardt	CIESIN - Columbia University, USA
Inês Lopes	Gulbenkian Institute, Portugal
João Luzio	Biblioteca Nacional, Portugal
Aurora Machado	Biblioteca Nacional, Portugal
Ana Canas Delgado Martins	Instituto dos Arquivos Nacionais, Portugal
Margarida Meira	Instituto Gulbenkian de Ciência, Portugal
Sam Pepler	NERC/CCLRC, UK
Flaminia Ramos	Fundação para a Ciência e a Tecnologia, Portugal
Diogo Reis	Biblioteca Nacional, Portugal
Fernanda Ribeiro	Universidade do Porto, Portugal
Robin Rice	Edinburgh University Data Library, UK
Carmen Rodríguez	Universidad de León, Spain
Seamus Ross	ERPANET, UK
Donald Sawyer	National Space Science Data Center, USA
Kevin Schürer	UK Data Archive, UK
Alexander Szalay	The Johns Hopkins University, USA
Jean-Pierre Teil	Archives nationales de France
Peter Thorpe	ICDDR,B: Bangladesh
Peter Van den Besselaar	Netherlands Social Science Data Archive (Steinmetz Archive) – NIWI, Netherlands
Peter Weiss	National Weather Service, USA
Henry Wolfinger	National Archives and Records Administration, USA
João David Zink	Biblioteca Nacional, Portugal
Thomas Zuercher Thrier	Swiss Federal Archives ARELDA, Switzerland

Appendix Three - Basic Definitions[*]

Access: the process of obtaining data from a storage device or system

Archival database: A database containing data values and other information retained over a period of time and represented as an accurate reflection of the contents at a specified time.

Archive: An organized and managed collection of information (in any form) that is protected to ensure its integrity as an authoritative source for the information stored in it.

Data: Scientific or technical measurements, values calculated there from, and observation or facts that can be represented by numbers, tables, graphs, models, text, or symbols that are used as a basis for reasoning or further calculation.

Database: A collection of interrelated data, often with controlled redundancy, organized according to a schema to serve one or more applications. The data are stored so that they can often be used by different programs with little or no restructuring or reorganization of the data. A systematic protocol is used to add new data or modify and retrieve existing data.

Database management. The activity associated with organizing, storing, and providing access to a computerized database. It usually includes responsibility for ensuring the integrity of the database.

Metadata. Data about data; consists of descriptors of data in a database to provide systematic information for users, application programs, and database management software. Metadata may be manipulated and searched and may themselves be organized in a database.

Raw data. Data as originally recorded and that have not been combined, modified, interpreted, or adjusted in any way.

[*] Source: National Research Council. 1996. *Bits of Power: Issues in the Global Access to Scientific Data,* National Academy Press, Washington, DC., and J.H. Westbrook and W. Grattidge, eds. (1991), "A Glossary of Terms Relating to Data, Data Capture, Data Manipulation, and Databases," *CODATA Bulletin* 23, Nos. I & 2, 196 pp.

Appendix Four – Sponsoring Organisations

CODATA (www.codata.org)

The Committee on Data for Science and Technology (CODATA) is an interdisciplinary scientific committee of the International Council for Science (ICSU), which works to improve the quality, reliability, management and accessibility of data of importance to all fields of science and technology. CODATA is a resource that provides scientists and engineers with access to international data activities for increased awareness, direct cooperation and new knowledge. CODATA was established in 1966 by ICSU to promote and encourage, on a worldwide basis, the compilation, evaluation and dissemination of reliable numerical data of importance to science and technology. Today 23 countries are members, and 14 International Scientific Unions have assigned liaison delegates. CODATA is concerned with all types of data resulting from experimental measurements, observations and calculations in every field of science and technology, including the physical sciences, biology, geology, astronomy, engineering, environmental science, ecology and others. Particular emphasis is given to data management problems common to different disciplines and to data used outside the field in which they were generated.

ERPANET (www.erpanet.org)

Increasing amounts of Europe's cultural and scientific heritage is being created or represented in digital form. The preservation and reuse of these digital assets forms both the cornerstone of future economic growth and development, and the foundation for the future of memory. This material represents Europe's heritage and is its future intellectual capital. The Electronic Resource Preservation and Access Network (ERPANET) widely recognises the benefits of using digital information and a result of its prevalence is the emerging vision of Europe as an information rich society whose record is just waiting to be harvested and processed by the technology-enabled researcher of the future or by emerging eContent industries. Ensuring this vision depends upon the survival of digital data in accessible and usable form.

The fast pace of change in the technological landscape makes ensuring technological advances preserving digital assets cannot happen as an after-thought, it needs to be planned. Policies, technical methods and strategies are required because media degrade (e.g. magnetic particles lose their properties and dye layers on optical media break down), technological developments make systems obsolete, or information is rendered inaccessible by changes in encoding formats.

ERPANET addresses the lack of awareness, fragmentation of knowledge and skills amongst the stakeholder communities about how to handle existing digital preservation problems, and how to plan effectively for the future. ERPANET tackles the lack of identification and focus on core research areas and brings coherence and consistency to activities in this area.

Biblioteca Nacional (www.bn.pt)

The BN – “Biblioteca Nacional” (National Library of Portugal) was created in 1796. Its role includes collecting and preserving the national bibliography, through the application of the Legal Deposit Law, to act as a standardisation institution in all matters concerning librarianship, provide access and disseminate information about its collections, and co-ordinate the National Union Catalogue (PORBASE). BN is also the National ISSN centre and the national representative for ISO/TC46.

In 1997 BN was assigned by the Portuguese government, in the following of the publication of the national “Green Book for the Information Society”, the mission to promote the development of the Digital Libraries. That has been done internally, by adapting its structures to the new requirements and projects, at a national level by participating and promoting working groups and events, and internationally by participating in projects.

The present role of BN is the result of an evolution and of its consequent adaptation to the communication and information characteristics of nowadays society. The main purpose is not only to provide the intellectual and scientific life of the country with all the cultural memory which its collections represent, but also to project its image abroad, thus playing an important role in the spreading of knowledge and fostering of modernity.