

## **Appendix C**

### **Detailed Information about Selected Projects**

#### **American Astrophysical Society - Peter Boyce**

We archive the scholarly journals in astronomy which are published by the American Astronomical Society. There is also an archive of all the core literature in astronomy from 1849 to the present (yes, it's 150 years, now) which we don't maintain ourselves, but which we collaborate with, linking the references and forward citations of the scanned articles with the new electronic versions of the journals and with each other. We also link to the international astronomical databases, so it is hard to tell where one stops and the next begins. I will answer for the AAS journals, but the importance is to maintain the linked access which my astronomical colleagues find so useful.

#### **Purpose**

The purpose is both to preserve, but more importantly, to maintain access to the material for use. This means the links have to work, the scripts have to work and the whole complex structure of our advanced electronic journal has to be preserved in working order. This is, of course, a prototype, but it is our working store of knowledge in astronomy.

#### **What is Archived?**

We archive the underlying master SGML copy of the journal which the public never sees, along with the scripts, programs artworkk, etc. which make up the working journal. The whole article in SGML and EPS graphics is archived. The original may contain electronic links to other material. Name resolution is used, not the URLs. The content of the link is archived only when it is an article from AAS journals, which many of the references are. They collaborate with the other information producers in astronomy to keep the whole astronomical information system current. There are three mirror sites around the world. Libraries save the paper, but that is no longer the complete journal, and is not sufficient. We do not specifically save paper except insofar as we have a complete copy of the paper journals.

#### **Refreshing**

They are refreshing for technical currently about once every 18 months. With the SGML master copy, this is a VERY small additional expense, and is covered in our operating budget.

#### **Format**

The native format for the input is LaTeX, MS Word or Word Perfect and EPS graphics. The native format is transformed immediately into SGML as the first step in the original publishing process. AAS uses its own translation software developed in Omnimark. In the future they will

translate into whatever new standard becomes accepted – such as XML, if that takes off.

## **Metadata**

Dublin core elements -- Author, Title, abstract, Author's institution, bibliographic data, etc. The metadata is at the article level for now. The identifier is Astronomy's standard identifier called a "Bibcode", which has been in use for fifteen years. We are adding PubRef numbers this spring, and we can add other identifiers within our information system.

## **Process**

The originator is the author of the manuscript. The archival process is one and the same as the publishing process. AAS's publisher, the University of Chicago press, is responsible for the storage and archival management of the material. The originator has no role in the archiving process.

There are no extraordinary backup and recovery procedures. There are no special agreements with the originators of the information for archiving, should AAS go out of business. Both AAS' journals are over 100 years old, and we think that we are more responsible than the creators of the material.

## **Technology**

Hard disks (RAID arrays) and optical platters are used. The technology migration plan is to keep up with progress in the technologies. This is a living journal. AAS expects to migrate as often as necessary. We build in new technology about every five years. With our process, the archive maintenance is indistinguishable from the general journal operation.

## **Costs**

Maintenance of the archive is much less than one percent of the overall journal budget. It is so small we have not split it out. We have established a fund which escrows a small part of the current income so that after five years we could do a major translation and migration of the journal. We have four years of experience now, and feel this is not necessary, because, under our process, the costs for maintenance of access and of the working journal are being well covered under current operations.

## **Policies**

The plan is to keep everything forever. Storage space is getting cheaper faster than our material is growing. No need to worry about retention and disposition under those circumstances.

## **Access**

The archive is accessible to subscribers via current browsers at any time for the cost of the current subscription. They may open the back issues after five years to anyone. That's under discussion. The NASA supported page image archive of the last 150 years of astronomical literature is absolutely free. The terms and conditions are the same as usual fair use and interlibrary loan conditions apply. Our license is liberal. The users can download the material to personal files. The format is PDF. The HTML material does them no good because the links are run through name resolvers to make them robust, and to make it possible to manage three mirror sites. Accessibility will be maintained over time by updating the material to make it readable by the most current browsers. The SGML remain unchanged -- until we go to a new standard such as XML.

## **Future Plans**

The challenges are maintaining graphical material and other non-textual information.

## **American Institute of Physics - John T. Scott**

### **Purpose**

AIP's product is a wide range of so-called "archival" journals in physics and related subjects. (The adjective predates the current interest in the electronic archive, and refers to the longstanding requirement by scholars for a body of literature that reliably records all published and established knowledge.) The electronic archive is a requirement for the continuation of this objective in the electronic era. The primary archive will not be used for routine document supply etc.; its goal is preservation. Copies in a secondary collection will be accessed by those with appropriate credentials. The archive is operational

### **What is Archived?**

Text and graphics files, with all bibliographic data and indexing, that make up the complete journal articles are being archived. Electronic links--to references cited in text, for example--are included (not the cited material itself). Auxiliary material (additional data or graphics supporting the article presentation, and made available on demand, but not published in the paper journal) is also archived. Paper is archived too, for those products that have a paper analog.

### **Refreshing**

It's too early yet to know how often "refreshing" will be necessary, but the files will be refreshed and migrated as needed.

### **Format**

Native format of text files is SGML, and it is retained. Plans for the future depend on what the future brings.

## **Process**

Some of the original material is owned by AIP itself (the AIP physics journals); some is owned by certain Member Societies of AIP--the society journals for which AIP provides a publication channel both on paper and online. The "originator" of the information is the article author, but the journal editor, practicing strict peer-review, is the source from whom AIP receives material for publication. AIP takes responsibility for archiving its own material, and provides an archiving service for those societies that request it. An offsite copy of the whole archive would provide a recovery channel in the case of loss of or damage to the prime archive.

## **Technology**

Migration is planned; frequency will depend on the rate of technological development in the industry.

## **Costs**

This information cannot be shared.

## **Policies**

See AIP's Website ([www.aip.org](http://www.aip.org)) for text of AIP's policy statement. Intellectual property concerns are as much a part of the electronic archive as they have been for the paper product.

## **Access**

The primary archive itself is not publically accessible (see above). Other accessible files are available via the WWW to journal subscribers. Journal articles may be downloaded to personal files, subject to the usual restrictions governing copyrighted works.

## **Future Plans**

The challenges are seen as coming from the inexorable growth of the journal-publishing endeavor. Costs in terms of both funds and staff time will rise year by year, and the annual charge will change from representing a small fraction of the total publishing cost to rivaling the costs of publishing the current volumes. Although AIP has been fairly successful in recent years in minimizing journal growth (to control subscription prices), there is a feeling among many publishers that online journals can be allowed to grow almost without limit. Archiving costs will be affected accordingly.

## **Atmospheric Radiation Monitoring Center - Ray McCord**

## **Purpose**

To store data and metadata generated by the Atmospheric Radiation Measurement (ARM) Program. Data from external sources that complement the Program's measurements are also stored. The Archive is active. It provides on going access and re-use. The quantity of data extracted from the Archive is primary performance measure by the sponsor. The Archive maintains the only copy of the data, so preservation is also an issue. Many of the data 'timeless, once-in-a-lifetime' measurements so preservation is also presumed. The archive is operational

## **What is Archived?**

All of the data generated by the program. This includes data, metadata, operations log, value added products, data quality reports, etc. The raw data are the first electronic signals captured by a data logger or dedicated PC on an instruments. The data includes measurements of sunlight, meteorology, clouds, temperature and water vapor profiles... several thousand measurements types. Most of the time all data are archived. A few instruments generate more data than can be retained (e.g., digital cloud radar and spectral image information). Slightly summarized data are retained for these.

The data do not contain links. Each file is nearly 'stand alone'. Some of the descriptive information is maintained on web pages. Eventually these will be integrated into a metadata database. The database will be cross-referenced by site, instrument type, and location. The currency is assured primarily by the migration to new technology. This migration is currently on 4-5 year cycle. Transfer to technology and media is a major effort. 99.999 percent of the ARM data originates in electronic form. The Archive has the only copy of the data that is dedicated to being retained and accessible. Other copies of some parts of the data may be located among the researchers, but they are not dedicated to being retained and available.

## **Format**

The format of the input varies; it is dependent on the output of the instrument and its data logging. The raw data is retained. The data are also transformed to a NetCDF format using the UCAR public domain software for handling NetCDF format. NetCDF is a widely accepted data structure used in atmospheric sciences. The data will be accessed with newer versions of NetCDF utilities. Some commercial data analysis software has access to this format as well (e.g., matlab, IDL).

## **Metadata**

Metadata include descriptions of sites, facilities, instruments and algorithms, measurements and operation conditions, time and date, and QA information. The level of the metadata varies with the type. Site descriptions may apply to a large domain of data. QA and observational notes may

reference only a very small part of the data.

## **Process**

The instruments in the field operated by the Program is the originator of the data. Some information is derived values generated by algorithms. The ARM data system contains many processes that interact with instruments and partition information into data files. All of the data are transferred by way of Internet connections. We receive new data files every day, approximately 2000 files per day or 6-8 GB.

DOE recognizes the need for sustainability should the current center no longer be funded., but has not yet dealt with this issue. The ARM program is expected to continue at least 5-7 more years. However, the usefulness of the data will continue 10-20 years beyond the end of the project. It is not know how the operational costs will be continued after the project (or how much of the data will permanently valuable.

## **Technology**

The storage technology is tape libraries; high capacity tape cartridges. Specifically tapes in 3490e and 3590 format. The libraries, tape drives, and media are developed by IBM and Storage Tek. We plan to migrate to new technologies every 4-5 years. During each migration, the data will be copied to a new technology. Because the Archive is a multiple Terabyte collection (currently 4.5 TB and 3.2 million files); each migration will require 6-12 months. This is major effort and may become nearly continuous as the size increases.

Backup and recovery are performed using tape cartridges in an automatic (robotic) tape library. To prevent media accidents, 2 copies of the data are stored on separate cartridges in the tape libraries. A 3rd copy of the data is recorded by another tape drive and stored in a separate building.

## **Costs**

The budget is \$1.5-\$2 M per year. Funding in this range covers 1994 - 1999. It is expected to continue until at least 2005. The project costs to-date have been approximately \$8-9M in operating costs and \$3.5M in capital equipment costs. The cost to set up initial operation is difficult to define. However, the cumulative cost to reach a performance level similar to the current pace was approximately \$3.5M in operations and \$3M in equipment.

## **Policies**

We have some internal 'guidelines' for data documentation, file format and data transfer. Generally, we don't have many formal policies. We follow the fully accessible and free distribution policies developed by the US Global Climate Change Research Program.

We follow the climate change research program policy of openness. A very small portion of the data has some proprietary restrictions. These are distributed with written access agreements between the user and the provider.

Retention and disposition decisions have not been addressed.

### **Access**

The higher level, formatted data files are accessible via a user interface. This interface is accessible as a web site on the Internet ([www.archive.arm.gov](http://www.archive.arm.gov)) via web browsers. It is free at any time to anyone with Internet access. They request acknowledgement of data source be included in publications; we request copies of the publications. These are informal and not enforced. All data processing for research is conducted on the user's own system. The data format has been very stable for the past several years and we have not had to change the data structure to maintain access by the software.

### **Future Plans**

Within the next 18-24 months we will enter the first major migration phase. We also need to restructure the database to ensure that it does not become cumbersome as it grows. We have many metadata attributes to collection and develop. The archive is expected to grow to ~50-70 TB

The challenges include maintaining a consistent rate of performance for the system as it continues to grow. Finding new ways to present the data and metadata to the user as the data collection increases in complexity and scale. Maintaining flexibility in the storage and access logic as we learn more about the changing demands for the data. This archive supports a research program and the activities associated with the data (generation, collection, transport, access, analysis) are constantly changing and expanding.

## **Carbon Dioxide Information Analysis Center (CDIAC) - Robert Cushman**

### **Purpose**

To provide current access to, and long-term stewardship of, key data bases related to global-change issues, in support of CDIAC's role as a national data center and World Data Center-A for Atmospheric Trace Gases. It is both an active and preservationist archive

### **What is Archived?**

Data files and associated documentation, as well as some non-data products (e.g., computer models, newsletters, annual reports, catalogs, bibliographies, topical reports). The raw data is environmental observations and estimates (e.g., measured values of atmospheric or oceanic

concentrations of greenhouse gases, land cover estimates, coastal data, carbon emission rates, weather measurements). Data types include text files, ASCII data files, graphics images, proprietary data files (SAS spreadsheet, ARC/INFO, et al.)

In many cases our online documents are cross-linked. Other links go to external locations. These are archived if they are our own product.

Some products are also stored as printed reports. In many, but not all cases, documents are printed, in limited number, at the same time they are posted online. Some users prefer printed documents. Other documents are produced online only, but can be printed by the user.

### **Format**

Data are received in a variety of formats (e.g., ASCII files, spreadsheet files, ARC/INFO files, SAS files, printed tables). The format may be retained or transformed depending on the specific product. Transformation may be accomplished using ASCII, SAS, Excel, Fortran, ARC/INFO, using database, GIS, text editor programs. We expect to always provide ASCII format; the others are subject to change depending on what software is used by us and the user community.

### **Metadata**

Our documentation typically includes the following information: Name/affiliation of contributors; Suggested citation; Background - scientific significance; Methods - instrumentation, experimental design (e.g., sampling frequency, replication, controls), calibration; Data management - data reduction, data rejection, gap filling; Quality-assurance; Data formatting - variable names & definitions, units, missing values, quality flags; Recommended uses & limitations for the data; Fortran & SAS input statements; Partial data listings, integrity checks for data transport; Reprints of relevant publications.

### **Process**

A variety of investigators around the world - primarily university researchers and government agency scientists create the raw data. Data and documentation arrive in various ways - traditional mail, email and attachments, ftp transfers, fax. For some products, material is received essentially once, although there is typically follow-up correspondence concerning data, quality assurance, and documentation. Other products, especially environmental monitoring products, require updates on an approximately annual basis.

CDIAC relieves the data contributor of those burdens, although some may wish to maintain their own archives, particularly for active research projects.

CDIAC staff scientists work on the data, using their own networked workstations and PCs, with

temporary file storage in working directories. When data bases are ready for public access, they are moved to appropriate locations in the online area of our servers. They are considered to be archived at that point. Some data bases are taken offline when replaced by more current versions, but they are still archived and backed up, and available on special request.

If CDIAC's World Data Center-A for Atmospheric Trace Gases were to cease operation, ICSU protocols specify that we would have to transfer our WDC data holdings to another WDC. CDIAC's non-WDC holdings would be handled on an ad-hoc basis; there are no set procedures.

## **Technology**

Disk storage is used, with digital linear tape backup.

We keep up with new developments in digital storage. When a new approach is well established, we migrate our archives. Typically, the capacity of the new technology vastly exceeds that of the old, and it is easy for the new storage devices to absorb all of the existing holdings. And this ensures that we do not have the problem of data that are irretrievable because the media have expired or the hardware is no longer operational. We expect migration as often as appropriate, given the pace of market penetration of new technologies.

## **Costs**

CDIAC's total annual budget is currently about \$2.5 million. CDIAC's budget in FY1982, its first year, was \$470,000. For the period FY 1982-1998, the cumulative total budget was about \$31 million.

## **Policies**

CDIAC's policy has always been to recognize the data contributors as authors of our data products, and we provide suggested citations for users to achieve this. CDIAC operates in compliance with ICSU-WDC and US Global Change Research Program policies on unrestricted access to data. All data products are retained until they are superseded by more current versions, or until the data contributors request that they not be distributed. Data products are updated as determined by the availability of more current data from their contributors.

## **Access**

All data holdings, when approved by the contributors for distribution, are accessible without restriction or charge, by any and all users. Most users employ web browsers or ftp software. On request, data and documentation are also sent via standard mail or fax, in printed format and on a variety of digital computer media.

## **Future Plans**

It is CDIAC's hope to provide value-added (quality-assurance and documentation) and archival functions for the most important global-change data bases. The universe of such data bases is always expanding. The challenges are identifying the most important global-change data bases and securing the funds to keep up with the work.

## **Defense Information Technology Testbed - Tammy Borkowski**

### **Purpose**

The purpose of the archive is to support military operations (prototype was demo'd in Bosnia) and feed training and lessons learned systems.

The archive can be accessed from the theater for 30 days. After that, it is currently shipped via tape to Leavenworth, where it is loaded into the MAAS system, a multimedia data warehouse. I've not seen the MAAS demo'd yet, but apparently you can easily extract video clips and insert into PowerPoint and other presentations. This multimedia archive will be used to create training and doctrinal materials under the Advanced Distance Learning Initiative. Another long term goal is to preserve imagery that lead up to important battle decisions.

This prototype is a proof of concept for full-life management of multimedia records within the operational environment and as a first step towards a virtual research library for Unmanned Aerial Vehicle(UAV) video in particular and multimedia imagery in general.

### **What is Archived?**

UAV records include 5 components (MPEG, JPEG, audio, text, and metadata). The proof of concept showed that these components could be collected, converted, linked, searched and managed as one record.

Actual UAV feeds are analog VHS video. They are converted to MPEG and the audio file of the remote pilots narration is added. The whole thing is encrypted and transmitted to Joint Analysis Center (JAC) in Molesworth, England (I think via satellite link). The data is in MPEG, JPEG, wav and txt formats. The whole item is stored as an object.

The object's metadata contains internal links to the video clip, still image summary shots, pilot narration and transcript. The content of the link is archived.

### **Refreshing**

We are archiving electronically on the MAAS system at Ft Leavenworth. And we may also be keeping the tapes. I don't see a need for refreshing.

### **Format**

The native format is analog VHS. The native format is transformed to MPEG video.

### **Metadata**

Core and unique metadata is used IAW DoD STD 5015.2 and NIMA Core Video Metadata Profile. Looking into fully automating metadata generation. The metadata is at the whole item level. The location identifier is the latitude/longitude.

### **Process**

The creator of the original material is Predator and Outrider UAV platform. There is a realtime link to JAC in Molesworth, England. Every 30 days a magnetic tape is sent to Ft Leavenworth.

JAC in Molesworth and Ft Leavenworth jointly. Effort funded by National Technology Alliance. The creator plays no role in the archiving function.

VHS video is recorded in theater by reconnaissance UAVs. From command post, remote pilot narrates mission. Video is converted to digital MPEG and audio is digitized (wav). Video and audio are encrypted and transmitted to JAC in Molesworth, England where value is added - audio file automatically transcribed, metadata generated, mosaic file(JPEG graphic that summarizes overall track) automatically generated, and mission profile file extracted from operator entered mission profile data. Looking into both push and pull technologies to get this information back to the commander in theater.

There are no extraordinary backup and recovery procedures. There are no arrangements with the originators or users for continuation of the archive if the organization is no longer in existence.

### **Technology**

MAAS currently being used. Also experimenting with using RetrievalWare to recognize video data in terms of allowing content-based querying

Plans will be in sync with C2 architecture migration. Also tracking DISA's Common Operating Environment and Joint Technical Architecture standards. Also using Standards Profile for Imagery Archiving (SPIA)

My personal opinion is that compression technologies will drive the technology migration, not storage or retrieval.

### **Costs**

To develop detailed specifications and install an unclassified video archival suite has cost \$22M. The start up costs were \$23M over 3 years. The projected ongoing costs are about \$300,000.

## **Policies**

This effort is part of the effort to come up with an Enterprise Records Mgmt solution for DoD and Federal agencies. Working with NARA on standards and processes. Also concerned with declassification efforts. Long term we will keep video online for 6 months and near online for another 6 months (i.e. it could be loaded/transmitted within a short period of time versus in realtime).

## **Access**

Ultimately the archive will be accessible to the commanders in theater through the integrated C2 systems (GCCS). The archive should be accessible as an operational system in 2005. Security issues, currently bandwidth issues, but expect that not to be a problem long term. Users can download the material to personal files at this time for testing and demo purposes. Not sure about this in the long term. They are working with NARA on maintaining access as software changes these issues.

## **Future Plans**

Create training/educational systems for warfighters and develop knowledge gateway connecting DoD databases. There is no maximum or optimal size to the archive.

The challenges include: Interoperability with C2 systems; Creating and following standards (technical, architectural, data models, metadata, etc.); Coordinating similar/competing efforts throughout DoD; Getting funding

## **Distributed Object Computation Testbed (DOCT)/San Diego Supercomputer Center - Reagan Moore**

It was too late in the data collection to obtain a completed detailed questionnaire from Dr. Reagan Moore at SDSC. Pasted in below is his e-mail message with links provided. The information in the report was taken from the content available at these web site.

Additional information can be found at <http://www.npaci.edu/DICE> under the NARA project. This includes a description of the NARA project, related publications, and presentations. We are adding the slides I used at the Workshop to the publications section of the NARA project. The project with the USPTO is described at <http://www.sdsc.edu/DOCT>

The approach we have taken can also be used with ecology data. We are working locally with John Helly, who is assembling an ecology database for ESA.

## **Electronic Publications Preservation Project/NLC Electronic Collection**

The most detailed information available is from the Web site at [collection.nlc-bnc.ca/coll-c/index-c.htm](http://collection.nlc-bnc.ca/coll-c/index-c.htm). See the description under the previous All Projects table.

## **EVA - Finnish National Library/Helsinki Univ. Library**

No detailed response was received. However, an English version of the EVA Technical Requirements and guidelines were received. The contact to receive a copy of these guidelines can be provided upon request.

## **HighWire Press**

HighWire was unable to respond to the detailed questionnaire due to preparations for an upcoming users/publishers conference. The information contained in this detailed outline is taken from the HighWire Web site.

### **Purpose**

The purpose of HighWire is to provide an avenue for small university and learned society publishers to produce electronic publications, particularly electronic journals. The archive is considered to be part of the services provided.

### **What is Archived?**

The primary archives are text and images, since many of the journals are still mirrors of their print versions. The HighWire archives begin with whatever the first issue is of the journal that they helped to produce electronically. However, HighWire is committed to creating and archiving more interactive, multimedia journals that include extensive links and connections to other types of data objects.

### **Format**

The material is accessible and archived as PDF.

### **Metadata**

The metadata available for searching is the standard bibliographic journal header record. However, with Adobe software the content of the article can be searched. There is no information available about special metadata elements for preservation.

### **Costs**

The costs vary by the price set for the journal between the society and HighWire. There is no cost of archiving available. Access is based on subscription. Site licenses are available.

## **Access**

Access is by the Web.

## **JSTOR**

JSTOR did not respond to the detailed questionnaire, but a comprehensive brochure was received from Kevin Guthrie, President of JSTOR. The information contained in this detailed outline is taken from that brochure and from numerous sources on the Web.

## **Purpose**

The purpose of the archive is to build a reliable and comprehensive archive of important scholarly journal literature, to improve dramatically access to these journals, to help fill gaps in existing library collections of journal backfiles, to address preservation issues such as mutilated pages and long-term deterioration of paper copy, to reduce long-term capital and operating costs of libraries associated with the storage and care of journal collections, to assist scholarly associations and publishers in making the transition to electronic modes of publication, and to study the impact of providing electronic access on the use of these scholarly materials.

JSTOR's purpose is to archive, preserve and provide ongoing access. Users can access JSTOR via the World Wide Web.

The project is not a prototype, but rather JSTOR has established itself as a non-profit agency.

## **What is archived?**

Phase One of JSTOR archived journals from 15 disciplines and includes 117 journals. Phase II will be a general science cluster that will include publications like Science. Future phases continue to focus on smaller disciplinary clusters.

The data being archived is derived from both hard copy publications and electronic media.

If by "whole" item you mean the entire issue of a publication, yes. In addition to articles and graphics, advertising material, membership lists and any other miscellaneous content are scanned. JSTOR scans the whole publication, cover to cover.

In a review of articles about JSTOR, I did not find any mention of electronic links to other material.

JSTOR only stores the digital copy of the publications.

## **Format**

The native format is the hard copy or the native format of the electronic journal, presented in ASCII. JSTOR provides only images to its users, but uses text files created by using OCR software to facilitate searches.

According to an article written by Margit Dementi, "Access and Archiving as a New Paradigm", in the Journal of Electronic Publishing (<http://www.press.umich.edu/jep/03-03/dementi.html>), JSTOR recognizes that there are problems surrounding long-term archiving of electronic materials, because of the constant evolution of technology, but admits that there is no single solution that can be employed today.

## **Metadata**

There was no mention of the term metadata in the articles I reviewed or the JSTOR web site.

## **Process**

Copies of journals that are being archived by JSTOR come from various sources such as the publisher, libraries and journal replacement services.

According to the JSTOR Production Process page on their web site (<http://www.jstor.org/about/production.html>), the production process is broken into several steps. Once the copies of titles are received from the sources mentioned in the above question, the JSTOR production staff inventories them to confirm that a complete run of the title exists and through a page-by-page examination of each issue, creates a publication record for the title. Preservation concerns are addressed during this part of the process and scanning guidelines are created. A serials specialist on the staff examines the structure of each title and creates appropriate indexing guidelines, matching the indexing specifications used by JSTOR to each individual title and article. Finally, each journal title is shipped to a contractor for scanning and data entry. At the contractor's facility, the journals are disbound and separated into discrete issue units. Each page is then scanned at 600-dpi resolution, with meticulous attention to quality. Page images are checked for marks, folds or skewing and are rejected if deemed unacceptable. Each page image is then processed by OCR software in order to create a digital file of textual information. Quality control operators at the contractor's site use spell-checking software, review the text file created by OCR and correct errors raising the accuracy of the text to 99.95%. A table of contents file that includes bibliographic citation information and an item type identifier as well as key words and abstracts if they exist in the original publication are keyed in for each article in the journal run. All three digital files created by the contractor, the page image, the OCR text file and the table of contents file are downloaded to CD-ROM for shipment back to the JSTOR production facility. At the JSTOR production facility, the files are uploaded from CD-ROM to the JSTOR file servers. Mirrored file servers are currently located at Princeton University, University of Michigan and in the UK at Manchester Computing Center. In terms of how often they receive their materials, JSTOR utilizes what they call a "moving wall". JSTOR starts with the very beginning of the journal's run and then lets the publisher tell them how far forward they can go.

For example, if the publisher specifies a moving wall of five years, JSTOR includes all issues up to five years ago, and move the journal forward in JSTOR one year at a time, while maintaining a five-year lag.

JSTOR has publishers agreements that specify that the publisher owns the copyright rights of the digitized images that JSTOR creates. Publishers are obligated to grant JSTOR perpetual license to the digitized archive that JSTOR creates. If the publisher withdraws from JSTOR, this perpetual license is limited to the right to provide the archive only to those libraries that were participating in JSTOR as of the date of the publisher terminates their agreement with JSTOR.

The JSTOR publication license is nonexclusive and allows publishers to license others to digitize the back issues of their journals.

JSTOR does not require publishers to commit to them for a fixed term, rather an agreement is set up that either party may cancel on six months notice. JSTOR feels that this agreement will keep them responsive to the publisher's needs as well as those of libraries and scholars.

#### Technology

This is not mentioned in any of the literature reviews. There was no response from JSTOR on this question.

#### Costs

The only mention of cost is from an article by Kevin Guthrie, entitled, "JSTOR: The Development of a Cost-Driven, Value-Based Pricing Model,": available at (<http://www.arl.org/scomm/scat/guthries.html>). In his article he mentions an estimated cost of \$2.5 million annually.

#### Policies

See answers under Process.

#### Access

The archive is accessible via the Web to members of subscribing institutions. There is no mention of specific software. Access is through current browser technology.

The service is available at varying fees to colleges, universities, foundations (both in the US and internationally) and to scholarly societies. According to Kevin Guthrie in his article, "JSTOR: From Project to Independent Organization," available at [www.dlib.org/dlib/july97/07guthrie.html](http://www.dlib.org/dlib/july97/07guthrie.html), the cost structure is based on the size of the institution. JSTOR uses The Carnegie Classification of Institutions of Higher Education for

pricing purposes because this grouping reflects an assessment of an institutions commitment to research.

According to the JSTOR web site ([www.jstor.org/about/](http://www.jstor.org/about/)), there are two types of payment: 1) a one-time Database Development Fee (DDF), for permanent access rights to information in the Phase I archive; and 2) an Annual Access Fee (AAF), to help cover the recurring costs of updating and maintaining the archive.

For International libraries, there is no equivalent to the Carnegie Classification for grouping academic institutions outside of the United States. JSTOR aims to match the contributions non-U.S. institutions make to the value they derive from participation. Through analysis of JSTOR usage and collecting patterns at participating libraries, they have developed a methodology for setting value-based fees for libraries around the world. Institutions are first placed into JSTOR classes ranging from Very Large to Very Small. Fee levels are then set taking into account the relative value of the JSTOR journal titles to the higher education community in the country as well as the local availability of fiscal and technological resources.

There was no pricing information for individuals or publisher participants included on the JSTOR web site.

According to the JSTOR web site, highlights of US libraries licenses are: Initial term of three years, with automatic one year renewal terms. Broad definition of "Authorized Users", which includes not only students, faculty and staff, but also anyone present in the library. Access is not limited to the library building. Authorization is generally controlled by using IP addresses. User Rules permit one printout as well as one electronic storage copy of any article or articles from the database, for the user's personal, noncommercial use. JSTOR will work with libraries to achieve stability and standardization. Licensee may use materials printed from JSTOR in interlibrary loan. Publisher contact information is provided in a publicly accessible area of the JSTOR website, so that requesting libraries can contact publishers directly, usually by email, for the materials sought through interlibrary loan. JSTOR provides customer service by e-mail phone and fax. Helpfiles and user documentation are also available online.

Hardware and software requirements: Macintosh, PC or UNIX workstations with Internet connectivity and TCP/IP installed; and Direct parallel or LAN-attached printer(s). A monitor with resolution of at least 800x600 is recommended for optimal performance. In addition, Netscape 4.0 Internet software is preferred over other browsers, and PostScript Level 2 printers offer highest speed printing when this option is selected. Internet connectivity of at least 1.5 mbits/sec data transfer capacity is helpful for fastest access.

### **Future Plans**

Future phases will focus on disciplinary clusters.  
The challenges include keeping up with emerging technologies.

## **Kulturaw3/Royal Library, National Library of Sweden**

The project manager was unable to respond to the request for detailed information. Most of the web site content is only in Swedish. The most complete information in English is available from the web site at [kulturaw3/kb/sc/html/projectdescription.html](http://kulturaw3/kb/sc/html/projectdescription.html) See table on All Projects for information.

## **Long Term Ecological Research (LTER) Network - John Porter**

### **Purpose**

Our goal is to promote ecological science by fostering the synergy of information systems and scientific research. See

[http://www.lternet.edu/documents/Reports/Data-management-committee/1995-DM-committee-report/im\\_1995\\_report.htm](http://www.lternet.edu/documents/Reports/Data-management-committee/1995-DM-committee-report/im_1995_report.htm) for a full vision statement.

We are charged with fulfilling both missions, providing both access and archival storage.

We have both fully-functional modules and prototype modules. Prototypes are aimed at providing a proof-of-concept prior to enhancement as a fully functional module. We are creating a Network Information System (NIS). The mission of the NIS working group is to design and develop a distributed, LTER-wide information system using a modular approach, while maintaining and building on present functionality

([http://www.lternet.edu/documents/Reports/Data-management-committee/1996-DM-committee-report/im\\_1996\\_report.htm](http://www.lternet.edu/documents/Reports/Data-management-committee/1996-DM-committee-report/im_1996_report.htm)).

### **What is Archived?**

Ecological data is archived by all sites in the LTER Network. Some sites also archive textual material in the form of proposals, theses, papers and research summaries. The LTER Data Catalog contains over 2,000 entries (<http://lternet.edu/DTOC>).

The LTER Network archives a broad range of data types. These vary from small datasets, such as analyses of soil cores, to large GIS data layers and remote-sensing data. Many different types of data are included. The most common archival format is as ASCII text, but information management systems at LTER sites use a wide range of approaches and deal with a large number of data types.

There is wide variability in the use of linked materials to dataset metadata at LTER sites. Where links are used they may be local (e.g., another part of the archive) or remote (with no archival storage). Metadata for many LTER datasets are generated dynamically from relational databases, although text copies may also be maintained. Although the majority of LTER sites use digital

archival media, some LTER sites use paper copies for archiving small but critical datasets. Often this is in conjunction with printed checksums that facilitate checking of optical character recognition

(<http://www.lter.umn.edu/tools/t1004.html>). The goal of these systems is to assure data preservation in a human- and machine-readable form over the next century or more.

## **Format**

LTER sites receive data in almost all possible formats! Common input formats are fixed column and delimited ASCII, spreadsheets and proprietary GIS and remote-sensing formats. See <http://www.lternet.edu/ecoinformatics/guide/baker2.fv2.htm> for software used at LTER sites (1992-1997). Sites vary in the degree to which they archive the native format or just transform the data to archival forms. Approaches to software changes take two forms. Some sites store data in ASCII files that require little updating to be suitable for future software. Other sites maintain the data in relational databases which can produce export files in a variety of forms. Both of these methods are robust with respect to changes in software.

## **Metadata**

LTER sites use a wide array of metadata standards at individual sites, many of them site-specific and dating back to the early 1980s. See *Data Management in Ecological Sciences*. 1986. William K. Michener - Editor. The Belle W. Baruch Library in Marine Science No. 16. University of South Carolina Press, Columbia, SC. for information on formats and procedures 1980-1986, *Environmental Information Management and Analysis: Ecosystem to Global Scales*. 1994. William K. Michener, James W. Brunt, and Susan G. Stafford - Editors. Taylor and Francis, London. 555 pages. for approaches 1986-1994 and <http://www.lternet.edu/ecoinformatics/guide/frame.htm> for information on LTER information systems 1994-1997.

However, the LTER Network also has specific content standards for metadata exchange ([http://www.lternet.edu/documents/Reports/Data-management-committee/1994-DM-committee-report/im\\_1994\\_report.pdf](http://www.lternet.edu/documents/Reports/Data-management-committee/1994-DM-committee-report/im_1994_report.pdf)), which overlap with many of the FGDC metadata elements.

Several LTER information managers were also involved in crafting the "Non-geospatial metadata for the ecological sciences. 1997. William K. Michener, James W. Brunt, John J. Helly, Thomas B. Kirchner, and Susan G. Stafford. *Ecological Applications* 7:330-342." standard.

See: [http://www.lternet.edu/documents/Reports/Data-and-information-management/IEEE\\_metadata\\_1997/LTER-metadata.htm](http://www.lternet.edu/documents/Reports/Data-and-information-management/IEEE_metadata_1997/LTER-metadata.htm) for a paper on LTER Metadata in 1997.

Note that these standards are undergoing revision and extension. We are actively pursuing technologies (e.g., XML, RDF, Z39.50) that will facilitate exchange and cross-walking with other

standards (e.g., FGDC).

## **Process**

The originator of the material being archived is typically an investigator at an LTER site or a technician employed by the site. Some sites also accept material from other, non-LTER affiliated, researchers or participate in federal databases (e.g., those sites at USFS Laboratories). The method of ingestion is site-specific, with some sites requiring the data provider to do the data input and other sites employing data entry technicians for data entry. Automated data loggers are also widely used.

Archival management of the material is typically the responsibility of each site's information manager, with the creator playing a relatively minor role.

The data flow varies between sites. The 1997 DIMES volume (<http://www.lternet.edu/ecoinformatics/guide/frame.htm>) provides profiles of the information management systems at a number of sites, as does "Global networks for environmental information: Proceedings of Eco-Informa '96. 1996 November 4-7; Buena Vista, FL. Ann Arbor, MI, Environmental Research Institute of Michigan (ERIM)"

Each site is expected to maintain off-site backups of all data. In the event that a site is discontinued, all data is transferred to the LTER Network Office. For example, the data from the now-discontinued North Inlet LTER site can be found at:  
<gopher://time.lternet.edu:70/11/catalog/data/nin>

## **Technology**

<http://www.lternet.edu/ecoinformatics/guide/baker2.fv2.htm> contains information on the hardware and software in use at LTER sites from 1992-1997. Each LTER site migrates individually as needs and opportunities dictate. At any given time at one or more LTER sites are undergoing significant upgrades.

## **Costs**

There is no specific budget for information management at LTER sites. Individual site research grants (varying from \$560K to \$1.2M, with most at \$700K) are responsible for supporting information management activities. It is estimated that 15% of site budgets go into information management activities. The LTER Network Office has additional responsibilities regarding network-wide activities and commits a higher proportion of its funds to information management.

## **Policies**

Guidelines for information management policies were established in 1993 and revised in 1998. See

<http://www.lternet.edu/research/data/netpolicy.html> for the current LTER-wide policy and <gopher://time.lternet.edu/00/doc/dgdl.txt> for the 1993 guidelines.

Development of site policies that balanced community access to data vs the rights of the data provider to first access to the data have been critical. See Porter, J.H. and J.T. Callahan. 1994. Circumventing a dilemma: historical approaches to data sharing in ecological research.

Pages 193-203 In W.K. Michener, S. Stafford and J.W. Brunt (eds.). "Environmental Information Management," Taylor and Francis, Bristol, PA for a detailed analysis of individual LTER site policies.

The policy listed above (<http://www.lternet.edu/research/data/netpolicy.html>) is a national policy for LTER sites.

Retention and disposition decisions are reached at the level of the individual site. For the most part, all data are retained.

### **Access**

LTER data are generally available via the WWW. The LTER Network server <http://www.lternet.edu> is the best starting point. Terms and conditions vary among sites and most sites provide immediate access to downloadable data without the use of specialized software. Almost all access is free to researchers and the general public.

### **Future Plans**

The information resources of the LTER Network continue to evolve, as they have since the network inception in 1980. Experience has taught us to use bifocals, focusing both on immediate challenges and long-term opportunities. The LTER archives will continue to grow at a fairly steady rate, although there was a dramatic increase in activity in 1994 when the LTER Coordinating Committee first mandated that site data be placed online.

There are three major challenges. The first is keeping up with the high rate of technological change. The second is working on developing a culture of data sharing within ecology that rewards data providers and data users alike. The final challenge is the need to continually improve systems to meet the heightened expectations of ecological researchers.

## **NASA Distributed Active Archive Center - Larry Voorhees**

### **Purpose**

The mission of the ORNL DAAC is to manage, archive, and distribute biogeochemical dynamics and ecological data in support of NASA-funded field studies, the global-change research community, policymakers, educators, and the general public. The ORNL DAAC is one of eight DAACs established by NASA as part of the Earth Observing System Data and Information System (EOSDIS). EOSDIS is part of NASA's contribution to the U. S. Global Change Research Program's effort to develop a predictive understanding of the global environment.

The ORNL DAAC is an active archive. The ORNL DAAC is an operational data center.

### **What is Archived?**

The ORNL DAAC archives and distributes biogeochemical dynamics data and associated documentation. This consists primarily of *insitu* measurements, with some remote sensing products as they relate to the field investigations. The vast majority of data are archived as ascii files. A relatively small number of geographic information system (GIS) and remote sensing data files, which may be in some other file formats. The goal of the ORNL DAAC is to assemble a comprehensive collection of data on ecosystem biogeochemistry and related environmental conditions, along with associated remote sensing data. All data and information are 'stand alone' products but may be related to other data available on the Internet. Occasionally, the ORNL DAAC includes links to other data and information in order to assist the user community, but all data and information in the archive are intended to be able to be used as 'stand alone' products. In cases where we link to other Web sites, we typically do not archive the contents of those Web sites. An example of where we do archive another Web site is where we 'mirror' a Web site from Brazil in order to provide better connectivity to this information for the North American user community. We also link to Web sites of NASA field investigations whose data will ultimately be archived at the ORNL DAAC (e.g., the Boreal Ecosystem Atmosphere Study; BOREAS); when the BOREAS Project is over; the contents of that BOREAS Web site (as well as the data files and documentation) will be transferred to the ORNL DAAC. We do not archive on paper or microfiche.

### **Format**

We receive the data in whatever formats the data generator wants to provide. Typically, the data are provided in ASCII format, and the data are archived and distributed in ASCII format. We also have geographic information system (GIS) and remote sensing data, which may be in some other file format. The format of the data sets must always be readable by the user, otherwise there would be no reason to keep it in the archive. Therefore, either the data formats will be changed to reflect the changes in technology, or software will be provided that can convert the data files into formats that are usable.

### **Metadata**

The ORNL DAAC, as well as all of EOSDIS, has adopted the Global Change Master Directory (GCMD) metadata standards. At present, there are approximately 5000 metadata descriptions of data sets in the GCMD data base from over 470 government agencies, research institutions, archives and universities worldwide.

The metadata reflect what is in the data set; one data set may have one or many data files.

All data sets have a unique identifier. They can be located by this identifier, if it is known by the user, or a user can search for a particular data set based on the metadata description of the data files.

## **Process**

The vast majority of the ORNL DAAC data holdings are generated by numerous principal investigators funded by NASA for an integrated field investigation. Other data sets are the result of gleaning the professional literature or have been compiled by other researchers throughout the world. Our data holdings are received following a field investigation or study, which occurs approximately every 2-3 years, and as they are compiled from the literature or provided by other global change researchers. The ORNL DAAC is responsible for archiving and distributing data to the user community. The data generator/provider is asked to provide as much documentation as possible about the data; however, we will work with whatever can be provided.

When we receive a data set, we prepare documentation about the data set and generate metadata that describe the data file. The metadata are ingested into a relational data base, which is used for locating the data file in the archive, typically as an ASCII file. This information is also provided to a system-wide data base so that a user can query the data holdings of all NASA DAACs from a single user interface. The DAAC's data holdings are also registered in the Global Change Master Directory.

## **Technology**

The DAAC's data holdings consist of 'tabular' *insitu* field measurements and remote sensing data. All of the *insitu* measurements are stored on spinning disk. Most of the much larger remote sensing data files are stored on tape in a robotic, near-line storage system.

Media are refreshed on a schedule that is dependent on the medium. We have copied data from more than 600 9-track tapes to tape cartridges. The ORNL DAAC routinely backs up all of its data and information.

Backups of the entire DAAC system are made on a regular basis. There are no firm plans for how our data holding would be maintained if the ORNL DAAC no longer existed, but NASA is currently discussing long-term 'inactive' archive' options with other federal agencies.

Migration is an evolutionary process that is constantly changing; i.e., we are migrating to new technologies as they prove themselves and we are able to afford to do so.

## **Costs**

Since starting in 1993, ORNL DAAC costs have been approximately \$13 million, or about \$2.2 million/year.

Start-up costs are difficult to determine, because so many factors are involved. In the case of a DAAC, however, it is reasonable to assume that approximately \$2 million is needed for start-up costs. DAACs are not stand alone data centers, but are part of a larger entity working to provide a 'one-stop shopping' environment for the user. This approach can cost more than a single data center because of the need to coordinate the entire process of metadata development, documentation standards, ingest, and a common user interface.

Our projected on-going costs are approximately \$2.5 million/year. This figure, of course, depends highly on the scope of work. The ORNL DAAC not only provides an archive and distribution function, but it also supports NASA field investigations and the development of data products.

## **Policies**

EOSDIS (i.e., the DAACs) has developed several policies, covering the following topics (not all of these policies are considered to be 'final'):

- Pricing and Billing
- Order Cancellation / Return / Refund
- Retention of Replaced Granules
- Informing the Scientists of Changes to Science Software
- Informing the DAACs of Inter-DAAC Data Delivery Delays
- Citation of EOSDIS Data
- Cross-DAAC Referrals
- Announcing Replacements of Data or Software
- Advertisement of "Bad" Granules
- Configuration Management in the EOSDIS Science Operations System
- Access to Data and Information Services
- Quality Assurance/Quality Control at DAACs
- International Distribution of Software
- Confidentiality of DAAC Users Information
- DAAC Role in Acceptance of System Releases
- Records Retention and Access
- Remotely Activating Test Production Plans
- Trouble Tickets, DAAC and Multi-DAAC

- Cross-DAAC Backup
- User Registration
- Orders by DAAC Personnel
- User Identification
- Approval for Advertisements
- Data Ingest
- Backup Recovery Exercises
- Archival of Special Products
- EOSDIS Data Format Standards
- Versions of Product Generation Executables

Intellectual property concerns have not been addressed, but we have participated in meetings with other national data centers on this topic.

NASA strives to make data available to everyone as soon as possible. However, the NASA Data Policy appears to be loosely interpreted, with no formal mechanism for enforcement.

The ‘Records Retention and Access’ policy in the above list of EOSDIS policies addresses DAAC usage records. This information is used for analysis of activity, judging the relative merits of DAAC access and distribution options, determining user satisfaction, responding to user requests and queries on older orders, and for predicting the future resource and system needs. Billing and accounting records are used for tracking financial transactions incurred in the processing of data orders. This information can be used to help determine when and if specific data sets should be turned over to a long-term inactive archive.

### **Access**

Data in the ORNL DAAC are available through a NASA System-wide Web interface, a DAAC-specific Web interface, a DAAC-specific FTP area, and through a User Services Office that responds to data and information requests via e-mail, phone, fax, and surface mail. The data are available through at least one of these mechanisms to anyone in the world, at no charge, at any time, 24 hours a day, every day of the year. The data are available at no charge, but we request that the user cite the data in any publication in which they may be used. How to cite the data is provided in the documentation with each data set, giving the data provider the credit and letting readers know where the data can be obtained. Data can be downloaded to personal files.

We maintain records of user request information and are able to notify users of any changes, errors, or updates to the data set. One of the reasons we archive and maintain the data in ASCII format is because that format is more universally acceptable for long-term use. If technology changes to the point of making our data unusable because of incompatible formats, then we will migrate to new technologies as they prove themselves and we are able to afford to do so.

### **Future Plans**

Future plans are to maintain the ORNL DAAC and to add more data to our holdings as resources allow. The ORNL DAAC has an advisory committee that provides scientific direction in prioritizing data sets to be archived as well as how to best serve the biogeochemical dynamics user community. The direction of the ORNL DAAC depends heavily of the user needs.

The primary challenge is to maintain a balance in the many competing demands for resources. Many times, when funding gets tight, data management and archiving of the data is the first thing dropped from a research program. Taking that approach is very short sighted. Environmental data are very unique . . . once a measurement is taken at a particular point in time and space, you can never capture that moment again because of the variability of nature.

## **National Digital Archive of Datasets (Public Record Office Project) - Kevin Ashley**

### **Purpose**

To permanently preserve public records from the UK which take the form of computer datasets (structured data of some form or another), catalogue them and document context and provenance, and provide access to that part of the material which is open to public access. The purpose is both to provide an active archive and in some cases an inactive archive. Some of the material is closed for various period from time from 5 to 100 years, for various reasons including including personal and commercial confidentiality. Some is open for access now, other parts will become available at various points, and some objects are partly accessible (i.e. available in an anonymised or aggregated form only for the present.)

This is a working public service. That doesn't meant we can't learn as we go along, but we've been doing digital archiving in one form or another for over 20 years so we hope we've got the basics sorted.

### **What is Archived?**

In broad terms, whatever we're asked to deal with by the PRO. Selection and appraisal is dealt with by them. We then deal with everything from that point onwards. We look for the dataset itself, any meta-data needed to reuse it (datatypes etc.) and contextual material in the form of paper or digital documents that explain its context and provenance - why it was created, who did it, how, who used it, when they used it, what they did with it, etc. We scan paper documents and place the digitised forms in the archive with the rest of the material. It comes from every part of Government, so covers a very wide range of topics potentially.

What we get and what we store are two different things. We have to take data in whatever format it originated in. We convert it to one of a number of standard flat-file formats with accompanying meta-data which describes structure and inter-relationships. The metadata performs multiple functions, generating catalogues automatically, controlling data access and display, and aiding in

migration and export.

There are various data types involved, but in the broadest terms, a mixture of database-like tables and digital documents.

The whole item is archived and more, in the sense that we seek out related contextual material that may have been created by others.

The original does not usually include electronic links to other material, but if it did, NDAD would have to maintain these links in some form or other. It's not a problem we've come across yet.

### **Technology**

The frequency is determined by a number of factors, including media age and number of accesses (we expect our media to have a working life of 5 years or 10,000 accesses, whichever comes first.) All data is monitored through periodic automated checks, with at least one copy being checked every six months. We hold multiple copies at multiple locations.

It is the only archival copy in most cases. For documents which originate on paper, we also hold the original paper document in archival conditions (controlled temperature and humidity, etc.), because the PRO want us to.

### **Format**

The native format can be absolutely anything. It is transformed, although we currently also hold bit-wise copies of the original material as well to allow our work to be repeated by others if necessary. This may not be a permanent arrangement (i.e. the original copies are not formally part of the archive, and there is no external access to them.)

The format into which it is transformed is one of NDAD's own devising. We could not find any extant standard that dealt with all the metadata issues we needed to deal with. Image files (from paper or digital documents) are, however, held in TIFF V6 format, using either G4 compression for bi-tonal images or RGB for full colour. The transformation uses primarily software developed by NDAD, although some conversion tasks are carried out either by public-domain software or commercial software (e.g. we use Microsoft Access to transfer Access databases into comma-delimited form and programs we wrote ourselves to transform metadata from it into our own metadata format.) Software changes will be handled by staying up to date with the technology through periodic reviews of the technical changes. We already are capable of exporting data in a number of formats from a single source format, and expect to use the same techniques to handle future migration without the need to interrupt access. This is a process we have carried out before.

### **Metadata**

'Metadata' can be taken to cover a large part of what we do. Micro-level metadata documents individual fields in a database, dealing with field names, data types access conditions and so on. Macro-metadata includes the contextual documents, information on data interrelationships, the catalogues which we produce, and the keywords with which we index the catalogues.

Metadata is at both the whole or part level. The catalogues conform to ISAD(G), an international standard for archival descriptions. This documents objects in a hierarchical form, from a series (which can encompass a set of datasets created over many years) down to the level of a piece, which is typically an individual document or column in a table.

A number of location identifiers are used. To the external world, a standard archival reference is a permanent reference which can be used to refer to a series down to an individual object in a way which will not change over time. Internally, we map these references to files, collections of files and pieces of files. This mapping process may change as technologies or software changes, but the external reference does not.

## **Process**

The creator or originator is the UK Government. Following notification from the PRO, we liaise with government departments to arrange transfer by whatever is the most convenient means. We deal with some tens of transfers per year (the number will increase as the archive grows older.) The ULCC is responsible for the storage and archival management. The creators must arrange transfer. They provide contextual information; they aid the PRO in the selection of material for preservation; they can comment on our catalogues (but do not have to.)

They have extraordinary backup and recovery procedures which guard against total physical destruction of the building and archive and the loss of all knowledge in the organisation. To this end, we hold off-site copies in a secure store, together with deposited copies of all our procedures and software to enable someone else to continue the work if necessary.

The arrangement for continuation is with the PRO (who are the legal custodian of the records, as the National Archive) rather than with originators. They are as permanent as any organisation can hope to be. Our relationship with them is a contractual one, and therefore must deal with issues relating to handover to another party at the end of the contract.

## **Technology**

We use a hierarchical storage management system, currently using D3 tape for nearline storage.

Our system design allows us to change any one aspect of the storage system (media, HSM software, access systems, data formats) without the need to change any other or to interrupt access. We cannot plan exactly when or what we will migrate to, but we have ensured that we will be able to do so with as little disruption as possible.

Migration schedules have varied enormously in the past. We expect the access software to change substantially at least once in the next five years; media and storage systems may have a lifetime of 5-10 years; data formats probably longer.

## **Costs**

We work to a predominantly fixed budget in our contract, with some provision for variable costs for higher levels of accessioning deposits. I'm not permitted to give details of the costs. The infrastructure was totally financed by our organisation, and we recover this investment during the lifetime of the contract.

Start-up costs for an e-archive may vary from \$2000 to many millions. It all depends what you need to do, how much material there is, how self-documenting it is, and the number of accesses and performance requirements.

They expect unit costs to fall as the size of the archive increases (many of the costs are fixed.)

## **Policies**

Our policies are partly inherited from the legal framework surrounding public records in the UK. But yes, it was important to establish clear agreement on mechanisms for deposit, performance monitoring and usage expectations in order to enable us to cost and plan the work.

Most of the material we hold is Crown Copyright, which is something I believe is UK-specific and applies to information created by government. Copyright of this form needs to be protected, but the considerations are different from those for more traditional material. Some of the material we hold is copyrighted by others, but its status as a public record means that access must be granted if there is no other reason (such as confidentiality) to prevent it.

By the time material reaches us, it has been selected for permanent retention. You can judge what this means by the fact that the PRO's oldest records are almost 1000 years old. More complex retention, appraisal and disposition schedules apply to material whilst it is still in the possession of government.

## **Access**

The archive is accessible, but not all its holdings are (explained above.) The catalogues are accessible almost without exception, even if the material they describe is not. A set of web interfaces was developed by ULCC.

Access is provided to anyone, anywhere in the world (for open material) or to the depositors and

designated individuals at the PRO (for closed material) access is free. Access requires registration, registration is personal to the holder (no institutional access) and material may only be copied for private study and research - no redistribution is allowed. Users can download the material only on payment of a fee, although we cannot stop anyone taking a copy by saving a screenshot. Watermarking allows us to trace copies, but is not foolproof.

Material is converted to avoid most of the proprietary software. But that is only one meaning of 'special'. We hold things like geographic information systems that definitely require special handling but they are all in formats defined by relevant standards.

### **Future Plans**

Immediate plans are to increase the holdings and widen awareness of the facility. We are also looking to possibilities such as creating educational access for schools (in the UK sense - i.e. those of 16 and under) using specifically developed resources as part of a UK-wide programme called the National Grid for Learning.

There is no maximum size in any sense that constrains us at the moment. We have 300 Terabytes of storage in the current system and can expand it 4 or 5 times without the need to change technology (although it would need funding.) We expect this to give us adequate reserve capacity for a number of years.

I could write a book about the challenges! In this case, relations with depositors are definitely foremost at present, since the deposition of computer data as opposed to paper records is a recent innovation for most government departments. Beyond that, developing access systems so that the material is comprehensible to everyone is the greatest challenge. I won't be happy until my Mum and Dad can use it easily (they're in their 70s if that gives you any idea of what I'm aiming for.)

### **Natural Environment Research Council (NERC) - George Darwell**

Our "archive" in the sense of the questionnaire is taken to include the infrastructure of data centres run by NERC for management of our scientific data. See our WWW page <http://www.nerc.ac.uk/>

### **Purpose**

Scientific measurements of the "environment" are unique, if only in the time of collection, and so in general are irreplaceable. As the lead body for environmental scientific research in the UK, the Natural Environment Research Council (NERC) pays great attention to the management of data collected as a result of its activities, recognizing these data as a key resource to enable and support future research. This is an active archive with on-going access and reuse. Preservation is implicit. The "archive" as defined above is operational, but evolving and improving continually, so we are always learning.

## **What is Archived?**

Mainly numerical scientific measurements of the environment. This includes some remotely sensed imagery from satellites and NERC's own aircraft. Data types are numerous; NERC covers the fields of atmospheric, oceanographic, ecological, hydrological, geoscientific and Earth-observation data, in the UK and frequently on a European or indeed Global scale.

## **Metadata**

Some metadata to our holdings exist already, and further activities are on-going in this area to improve our catalogues and their accessibility to outsiders.

## **Process**

The creators are scientists in NERC's own Centres and Surveys, or supported by NERC in UK academia. Other bodies depositing data with NERC either voluntarily or under statutory obligation. The process for receiving material is planned in the context of individual scientific projects. The particular NERC data centre concerned with that area is responsible for the storage and archival management. There are 7 designated data centres. The creator is responsible for liaising with the data centres, and agreeing to mutually acceptable data management plans.

Underlying back-up and recovery accords to appropriate best practice for scientific data processing.

## **Technology**

Various storage technologies are used by the centres. The plans for migration also vary. These are included in the long term plans of the individual data centres. As a rule of thumb technologies (storage/software) cannot be expected to have a life of much more than a decade.

## **Costs**

NERC is estimated to spend about 3million sterling per annum on scientific data management. NERC's total income from UK Government and from commissioned research etc. is in excess of 200 million. Since the environmental sciences are very data-intensive, much of our research is directly or indirectly data collection. The "archive" is a core NERC activity, and hence will continue, at least some level of funding, for the foreseeable future.

## **Policies**

NERC's data policy - downloadable from the WWW - sets out the principles governing the relationships between scientists collecting data, specialist data managers, and data users. Ownership of intellectual property may vary according to those who collected the data, their

relationship to their parent body, and contractual arrangements. IP rights are of course respected. Some scientific data are deposited with NERC by outside organizations under statutory obligations, and further data are donated. The majority are collected under NERC's own aegis. Retention and disposition schedules vary according to circumstances. In view of the irreplaceability of much environmental data, there is a presumption in favor of retention.

### **Access**

The access to the data is through specific software for that data. In general it is available to any applicant. Some data are embargoed until those who have collected them have had time to publish their findings. Some data are supplied in near-real-time [e.g. enhanced satellite imagery to research vessels at sea] although this tends to be the exception. Data are supplied on inexpensive terms for bona fide academic research. Commercial users are charged appropriately. Use may be restricted by the nature of the licence taken out, since commercial use will be precluded when the licence is on academic terms. Licence conditions are too detailed to be reproduced here, but will invariably preclude the unauthorized passing on of the data to unauthorized third parties. Some datasets are accessible on-line; others by transfer media. If access to the data requires special software, migration of this software to continue access is an integral part of the migration plans.

### **Future Plans**

Continued evolution of data services, taking advantage of new technology as it emerges, and tackling past backlogs as far as resources permit. As an integral component of scientific research, data management (like science itself) will always be constrained by the resources available. The requirement for new data services is likely to emerge as science itself moves forward.

The challenges include managing ever larger quantities of data, driven by novel data collection technology; the opportunities presented by e-commerce; the conversion of raw data understandable only by specialists into information of value to a wider range of intermediate- and end- users.

## **OCLC Electronic Journals Project - John Hearty**

### **Purpose**

Electronic journal program evolved from one electronic journal -- The Journal of Clinical Trials, and has evolved into a program where OCLC takes publishers data and makes it available to libraries on a subscription basis. OCLC sees this as a service to its member libraries, rather than a service to the publishers.

### **What is Archived?**

Electronic journal are archived. Some have print equivalents. OCLC has no project to digitize

back issues of paper journals; however, the publishers may choose to do this. There are currently arrangements for over 2200 journals from more than 46 publishers. 1500 are mounted at this point. The others are in process. They have been doing this for 2.5 years.

## **Format**

OCLC is only dealing with PDF. That is what is archived. There is no document delivery at this point. Haven't heard anyone being concerned about Adobe going away. They have ability to deal with other formats. Not sure if they are storing in SGML (perhaps for header info). Still not conducive to mass production.

We know that things are going to evolve. They have committed to adjusting to and dealing with the evolution, so that they will always have access to the journal.

## **Metadata**

Metadata is the most important part of the archive. OCLC is committed to working with international standards.

Dublin Core will be the fallback metadata format. It is being taken through the right standards process. They are experimenting with other pointer files. There does not need to be one metadata format (as their work with numerous secondary databases has shown). But there must be a way to search across the formats. The ECO pointer file is an A&I database. It will take a while to define an acceptable standard internationally. Crosswalks are very important. SICI codes are being used in some cases. DOIs are not currently being used, because the concept is that they will be loaded somewhere else.

## **Cost**

Access is available to OCLC member libraries on a subscription basis. The subscription is with the publisher or the jobber. They will also manage the subscription if requested by the publisher to do so. EBSCO, and others are doing this too. What differentiates OCLC is a commitment to the libraries that they will archive these journals and make them available to them forever.

If they go out of the business they will be able to take the copy themselves. The publishers must agree to this arrangement. On occasion they will link to the current issues at the publishers site, but a copy is still transmitted for archive purposes to OCLC. The primary reason for linking to publisher sites is that some publishers have their own online systems, and they don't want to distribute otherwise. The pointer file is in the bibliographic data. They come through ECO and this points them off. However, most publishers are now saying that OCLC should load the data because as an online utility they are more accustomed to handling high numbers of access and multiple simultaneous users.

The library subscribes to a journal, OCLC gives them access to the journal and keeps the database of the years that the particular library subscribed. The libraries pay a small access fee.

OCLC is currently working on the business and economic models for sustaining this service. Libraries look to OCLC more and more as a utility where they should be doing archiving for them. Over the summer parts of the report will be made available to the User Council and the Board. They do not feel that the current service will need to be used to pay for the archiving. There may be decisions made on a case by case basis that will allow access control and authentication to occur and that will allow for special discounts to certain groups.

### **Access**

Today you come in through an ECO looking Web-based interface. In the next 4 months will move this under FirstSearch 5.0 which will integrate ECO and FirstSearch. ECO pointer file is now available on FirstSearch. The efficiency of the system and other secondary databases would be linked as well. For example, a search on one of the secondary databases loaded under FirstSearch would show your libraries electronic journal subscriptions alongside the hit list. Of the 85 databases available under FirstSearch, those that are relevant will be linked to the actual journals. By profiling will be able to tell who owns what. At some point in time would like to do the document ordering component as well.

OCLC's involvement in the GPO ERIC digital library project was a pilot for this type of system. This pilot was completed at the end of December, 1998; the results are being analyzed. Technically, the electronic journals can be accessed from near-line storage devices at approximately a 10th of a cent per megabyte.

### **Future Plans**

Navigation within the item. Maybe create automatic TOCs for navigation, using TIFF images and wrapping PDF around it so that you have some navigation and links.

## **OhioLINK Electronic Journal Center - Tom Sanville**

### **Purpose**

The Electronic Journal Center (EJC) is OhioLINK's self-operated, multi-publisher, aggregated collection of electronic journals. After analysis of our options, the OhioLINK community determined that our own site would give us the best combination of performance, functionality, and integration with other resources and the archive. Our primary goal is the day-to-day expanded use of scholarly journal materials.

We typically license the complete electronic journal collections of publishers and make these available to our entire Ohio higher education community. Over time we will learn to what extent

this expanded access will result in expanded use. The initial data indicates such expansion is quite significant.

We currently have the collections of Elsevier Science, Academic Press, and Project Muse loaded. Upcoming loads will include APS, Kluwer, Wiley, and Springer-Verlag

Strictly speaking, it is inaccurate to think of the EJC primarily as an archive. The EJC is an access system that as a by-product creates an archive.

### **What is Archived?**

We receive and load bibliographic, table of contents, article abstract data and article full text from all current publishers. For APS, we will be only loading meta data and then linking to the full text on the APS home site.

Links to other resources are only supplied once the full data is loaded locally. The links are not archived and exist in an “on-the-fly” environment where links are verified upon the user selecting the link.

### **Refreshing**

The data is supplied one time by the publisher at the time of publication. Updates or error correction may later be supplied by the publisher, but the data is considered generally in “archive form” upon receipt. Regular tape backups are made in case of catastrophe, and the disks are RAIDed for further protection. No other copies of the data are made except by the end user for personal use.

### **Format**

The raw data, used for indexing, comes in ASCII. The image files are PDF images or HTML pages. Data is delivered either by CD ROM or by downloading it directly from the publisher’s site.

The native format is ASCII text used for indexing. PDF or HTML for images. The native format is retained. The ASCII is only used for indexing purposes and is not displayed. We see no immediate changes to format.

### **Metadata**

The bibliographic citation information including abstract is supplied by the publisher.

### **Process**

The originator is the publisher. We receive initial back file data on CD's or by downloading from the publisher site. Ongoing data is downloaded or received on CD ROM on a weekly basis. OhioLINK staff is responsible for the storage and maintenance. The publisher have no responsibility beyond supplying the data. The data flow is irregular for some publishers but improving. Most data is now received on a timetable similar to the publication of the paper version.

Backup of the full data occurs on a regular basis. The system is also configured for RAID 5 with a hot-swappable function that allows disks to be swapped without taking the system down. Plans are being considered for near- line storage on a robotic tape system.

We are not contemplating the OhioLINK program being incapable of continuing the archive function, so there is no backup archive identified.

### **Technology**

We are using Compaq Storage Works with RAID 5. Plans are for the use of near-line storage to be used especially for data that is not being heavily used.

### **Costs**

We consider this information confidential and also difficult to answer as a separate activity. The computer access and storage systems of the EJC are part of a larger complex of database services. The technical components described in this document are and will be increasingly used fluidly with other services. As noted in the Technology section we have made a considerable long-term investment in a multi-processor access system, significant disk storage, and the long-term ability to utilize significant, lower cost tape storage. The funds to pay for the EJC system are centrally funded as a part of OhioLINK's overall funding provided by the Ohio Board of Regents.

### **Policies**

There are no particular policies to support this archive. The development of the EJC is a part of the OhioLINK program's evolution based on established expectations. We have established funding models for each publisher that allow for the purchase of each publishers material for inclusion in the EJC.

All data added to the EJC is done so under license with publishers. The license defines our perpetual ownership and usage rights.

There are no retention and disposition schedules. We intend to archive and provide access indefinitely. The only question will be whether particular data is stored on disk or tape.

## **Access**

We provide access via software licensed from Science Server LLC. Access is provided to students, faculty, and staff at Ohio higher education institutions. There is no cost to users for access or downloading of articles. If a user chooses to print an article at a library-based printer there may be a printing fee (e.g. 10 cents per page). Use is for research and educational purposes. Users can download the material to personal files.

It is the obligation of the OhioLINK program to make necessary evolutions in the software and hardware platforms to maintain accessibility as an ongoing, funded program of the State of Ohio.

## **Future Plans**

We intend to continue the expansion of publishers included in the EJC; improve the computer architecture of our central site to maximize utility of disk and tape storage among our services, including the EJC; and the utility of EJC data through links to and from related data. We do not see that the likely growth in desirable data will exceed our ability to manage it.

You have to be willing to make the investment in infrastructure and have a purpose beyond archiving as a justification to doing this.

## **Preserving and Accessing Networked Documentary Resources in Australia (PANDORA) - Margaret Phillips**

### **Purpose**

PANDORA was initially set up by the National Library of Australia as a proof-of-concept archive to (a) test the business principles that were being developed by the PANDORA Project (b) obtain hands on experience with Internet publications and assess technical resources and capability. It was intended, however, right from the beginning, that the Archive would become an ongoing, long term archive, meeting the responsibility of the National Library to ensure long term access to Australian online publications.

Another purpose of the Archive is to establish a model that can be followed by other Australian deposit libraries which also have responsibility for preserving portions of the nation's documentary heritage.

The main goal is preservation - long term access to Australian online publications. However, we consider current access to be important too, and one of our business principles is to negotiate with publishers the right to provide networked access to titles in the archive. If permission to provide access to a title is not forthcoming, we do not archive it.

While the Archive was begun as a prototype, we now consider it to be an operational production

archive which will continue indefinitely. As explained above, we wanted to develop practical experience with archiving Internet publications, and to develop expertise, especially technical expertise, with all aspects of identifying, selecting, cataloguing, negotiating with publishers, capturing, storing, providing access to and preserving them.

### **What is Archived?**

Australian Internet publications are archived on a selective basis, according to selection guidelines that have been formulated by the National Library. Please see Guidelines for the Selection of Online Australian Publications Intended for Preservation by the National Library of Australia at <http://www.nla.gov.au/scoap/guidelines.html>

Scholarly publications of national significance and those of current and long term research value in their own right are archived comprehensively. Others are archived on a selective basis to provide a broad cultural snapshot of how Australians are using the Internet to disseminate information, express opinions, lobby, and publish their creative work.

The National Library considers that it alone cannot manage to adequately preserve all of the titles significant to the Australian documentary heritage. The task is too resource intensive and therefore costly. The National Library is negotiating with the State libraries (deposit libraries) to join it in the task, whereby the PANDORA Project would become the National Collection of Australian Electronic Publications. We have begun to develop collecting agreements that, if accepted by all of the State libraries, would result in both a broad and deep coverage of the Australia publishing output online.

Even with the participation of the State libraries, it is still envisaged that the resulting National Collection would be selective and that publications of low quality and low research value would not be included. It should be noted that while all titles that meet the selection guidelines are selected for archiving, some cannot be archived for technical reasons. For instance, pages that depend on programs elsewhere on the publisher's server, including pages that are created 'on the fly', have not been archived successfully to date.

The raw data are the files comprising the publications on the publishers' sites that are selected for archiving. The data types are the full range of formats used by publishers on the Internet including text (mostly HTML), images (gif, tif, jpeg), video, including streaming video, java scripts, audio files (e.g. .ra .mid .au. wav); pdf; and, style sheets (.css).

The answer to whether the whole or part of an item is archived, depends on how 'whole item' is interpreted. When a title is selected for archiving, we define the boundaries of the item that we want to archive. For example, the item might be a report or an e-journal that is part of a much larger university site. We do not necessarily want to archive the entire site and therefore specify to the gathering software that it confine its capturing process to the URLs relating to the report or e-journal only. Links back out to the larger university site are 'denied'.

In addition, the report or e-journal may contain links to external sites, that is, sites other than the university site. The gathering software is programmed not to gather these links. If these linked documents meet the selection guidelines in their own right, permission to archive them would

be negotiated with the publisher and they would be archived separately.

Links internal to the item are gathered. The Harvest Web indexing software is programmed to gather all files in directories subsidiary to the title level URL. For example, the URL for the e-journal Australian Humanities Review is <http://www.lib.latrobe.edu.au/AHR/>

The gathering software is instructed to gather this URL and all the files in subdirectories of AHR eg. <http://www.lib.latrobe.edu.au/AHR/current.html>

<http://www.lib.latrobe.edu.au/AHR/archive/Issue-December-1998/castro.html>

<http://www.lib.latrobe.edu.au/AHR/emuse/castro/mcauliffe2.html>

However, it is programmed not to gather anything above the AHR directory, eg

<http://www.lib.latrobe.edu.au/>

or any external site which may be linked to from it, eg

<http://www.otheredge.com.au/aa/castro/home.html>

The gathering software is permitted to gather to as many levels as necessary in order to gather the title fully.

The site is often “regathered”. Each title that we select for archiving is allocated one of the following gathering schedules: one-off, weekly, monthly, quarterly, half yearly, nine-monthly, annually. Factors including the importance of the information, the stability of the site, and how frequently the site is updated determine the frequency of gathering. A completed document that will never be added to is gathered once only as is some less important material that is gathered as an exemplar only.

This is the only archival copy. If the publisher offers print or microform versions in addition to the electronic version, we do not attempt to archive the electronic version but collect and preserve the print or microform version only (unless there are substantial differences between the formats). We do not ourselves create a print or a microform version of an electronic publication for preservation or any other purpose. If we archive an electronic title, then that is our only copy. We do not archive the electronic version if there is a paper or microform version because electronic archiving is resource intensive and expensive in terms of staff and technical resources and we have chosen to limit the archive to titles that cannot be preserved in any other way.

In addition, there are as yet no proven methods for the preservation of electronic files. We do know that we can preserve the print/microform versions and so prefer to collect them for preservation.

We do not create paper or microform copies of the titles that we archive because, in most cases, those formats would not adequately preserve the 'look and feel', the functionality or the file structure of the titles archived.

## **Format**

The native format is whatever the publisher has used on his/her site. 'Strictly speaking the native format is what is delivered via the web - it may be the publisher has an application/server side scripting system that delivers html - we archive the html that is delivered not the scripts etc. on the publisher's server.'

We have made the decision always to retain one copy of each title (and each version of a title, i.e., each gathering of a title) in its native format. We will make additional copies and transform/migrate them to the extent that is necessary to continue to be able to display a title for access. As changes to software and technical platforms take place, we will have to transform (migrate) files or adopt other preservation strategies.

We do not transform the native format to a standard archive format at point of capture, if this is what you mean.

We will transform to future formats to be able to maintain access to titles. For instance, changes introduced to HTML version 4 mean that deprecated tags used in earlier versions of HTML may not be supported by browsers compliant with HTML4. We will need to migrate/transform affected titles so that they remain accessible with full 'look and feel'. What software we use will depend on the particular situation in hand. In the example used in the previous question, we may be able to use validation software. Or we may need to write a small in-house program to identify files containing dead tags and to change them.

We will analyse each situation involving change of software, format or technical platform as it arises and develop specific strategies.

## **Metadata**

Each title that is archived is described at the whole of title level using a MARC record in the National Bibliographic Database. In addition, at this stage, a very basic set of administrative metadata is also recorded for each title, e.g., when the title was selected, when permission to archive was received, when the title was archived, what gathering software was used.

Because our technical set-up is still very unsophisticated, we are unable to record effectively the

full range of metadata that has been identified as necessary in our Logical Data Model. See <http://www.nla.gov.au/pandora/ldmv2.html>. We expect that the technical infrastructure that will be implemented as a result of our Digital Services Project Information Paper (see <http://www.nla.gov.au/dsp/>) will enable us to record the full range of resource discovery, administrative, preservation and rights management metadata.

At present, the metadata recorded is at the whole of title level. As part of the publication itself, we capture any metadata that the publisher may embed in the item and this may include metadata for parts of the item.

The Logical Data Model recognises that metadata for versions of a title, as well as parts of a title or version of a title will be necessary.

The DSP Information Paper specifies requirements for metadata that will index the content of the National Collection of Australian Electronic Publications, i.e., at part level.

The PANDORA Archive currently uses PURLs to link from the catalogue record (856 field) to the title in the archive. (We have used the OCLC software to set up a PURL generator on the National Library's Web site.) The various files associated with a title in the archive have a system-generated file name based on the URL or PURL from the publisher's site.

The National Library will be participating in a meeting of a working group of the Conference of the Directors of National Libraries in Washington in April to investigate a more effective system of permanent naming.

## **Process**

In most cases, we use a gathering robot to go out to specified URLs on the publisher's site and copy the files to the Library's server. The software that we use most is modified Harvest software developed by the University of Colorado. We have developed a program that interfaces with it to run the gathering schedules described under Process. The Harvest software does not successfully manage all archiving situations, and we also use WebZip to supplement it. In a small number of cases, publishers 'push' their titles to us, via FTP, or on a physical carrier such as CD or Zip Disk.

The National Library of Australia is responsible for storing and managing the titles that we are archiving. In addition, we are storing and managing a small number of titles being selected for archiving by the State Library of Victoria. Although ultimately it is envisaged that the State libraries will store and manage their own distributed archives, the National Library is undertaking this service in the short term while there are technical and resource obstacles to be surmounted.

The role of the creator/originator is usually limited to granting permission to archive. However, in a minority of cases, the creator/originator may be involved in assisting us to solve problems relating to archiving his or her site. For instance, one creator changed the structure of the file

names to facilitate archiving.

Some publishers 'push' their titles to us.

## **Technology**

Archive material is stored on a standard UNIX file system using conventional SCSI hard disk technology. Tape back up is to DLT tapes that are held offsite.

We will analyse each situation as it arises and devise a specific migration plan. We would migrate as necessary, as each new situation arose requiring a change in software, format or technology platform, to continue to be able to display archived items and retain as far as possible their look and feel.

## **Costs**

The staff in the Electronic Unit who build and manage the PANDORA archive are part of the Technical Services Branch. The National Library has made no additional allocation of resources in terms of human or financial resources to the PANDORA Project. The staffing (currently five staff) for this initiative has come from the recurrent funding of the Technical Services Branch. The cost in terms of staffing on an annual basis for the Electronic Unit is approximately \$300,000. However, the Library has allocated a sum of money from its recurring budget for the acquisition of software and hardware arising from the Digital Services Project. This will provide the PANDORA Project with a robust hardware and software platform for the future development of the PANDORA Archive.

In terms of the staff salary costs for the largely developmental work of the Electronic Unit and assistance from Information Technology staff and others, the project has cost approximately AUS\$1m since its inception in 1996.

Start-up costs have not been calculated as such. This depends very much on the actual nature of the e-archive.

Projected ongoing costs are not known in terms of hardware and software at this stage. This will only be known following the results of the Digital Services Project. The on-going costs in terms of staff costs will probably continue at around \$300,000-400,000 per year.

## **Policies**

We established business principles or policies to determine what we want to achieve and how we want to achieve it. Intellectual property concerns were addressed by seeking permission for use.

All of the titles so far archived are freely available on the Internet and publishers are happy for

users to access the archive on this basis. For each title that includes a copyright statement, we provide a direct link to it from the title entry page for the title, e.g., <http://www.nla.gov.au/nla/pandora/ahr.html> We also include a link to a copyright warning from every title entry page. To date, there is very little commercial publishing on the Internet in Australia. We expect that commercial publishers will be more concerned about intellectual property rights and realising their commercial investment. We are in the process of developing a Voluntary Deposit Deed which would specify the conditions under which the Library may provide access to commercial titles. In consultation with the publisher we envisage setting a period of time during which access to the title would be restricted to users within the Library building. Once the period of commercial viability has passed, we would want to be able to provide networked access to it free of charge.

The PANDORA Project and our model for a National Collection of Australian Electronic Publications reflects the National Library's (and State libraries') responsibility for the preservation of the national publishing output. The Commonwealth of Australia and each State has legal deposit legislation requiring publishers to deposit their works with the National Library and the relevant State library. However, the Commonwealth legal deposit legislation does not yet include provision for electronic publications. It is in the process of being reviewed. Some of the States have provision for electronic publications within legal deposit legislation and some do not.

At this stage, it is expected that all titles archived will be maintained in the archive in perpetuity. We recognise that at some future point, especially in the face of an expensive migration program, for example, decisions may be made to delete some material that time has shown to be less valuable. However, there are no such guidelines at present.

## **Access**

One title only has restrictions on a few files, requiring password access. Permission to access these files would be required from the creator. User access is via a standard Web browser. On the Archive side there is custom software associated with a standard Web server that delivers archived content.

There are no particular terms and conditions, although it is expected that users will respect copyright.

If they have the facility to download material from the Internet, they would be able to download files from the PANDORA Archive.

Access to titles that contain/require special software, plug-ins etc. may be able to be maintained. will not be able to maintain access to every one of these. However, wherever possible, the Library will endeavour to migrate standard formats from old to new versions in order to maintain accessibility. We will also investigate other solutions such as emulation where the value of the information would warrant the cost.

## **Future Plans**

The next step is to evaluate the responses to the Digital Services Project Information Paper and to make decisions about implementing a sophisticated technical infrastructure that will enable the Library to collect, store, manage, provide access to and preserve Australian Internet publications in an effective manner, and on a large scale.

The future plans are to draw the State libraries into the model as each becomes able to participate in the National Collection of Australian Electronic Publications.

We also want to implement a more suitable permanent naming system and anticipate doing so in conjunction with services to publishers that would encourage the generation of metadata as well as permanent names. We have been thinking about ways of creaming off this metadata to create a national metadata repository (centralised or distributed) which, in conjunction with a suitable search engine, could assist with resource discovery at the part of title level. We have also begun to discuss with traditional indexing services the possibility of cooperation to ensure that the citation contains the permanent name for the archived version, to ensure the longevity of citations.

There is no maximum size, because of the nature of the Archive, we expect it to keep growing indefinitely. Any implementation of a technical infrastructure to manage the archive will most likely have an optimum or maximum size, but that is a separate issue from the size of the Archive itself.

The challenges are legion and many pages could be written on this question alone. The principle challenges relate to acquiring or developing a technical infrastructure to manage a long term archive of Internet publications, and to finding the funds to purchase it. The preservation aspect is also a great challenge - finding ways to ensure that the titles archived are readable and accessible in the future. There are challenges ahead in negotiating satisfactory arrangements for all parties concerned (publishers, users, and libraries) in relation to access to commercial titles. Related to this is the need to develop ways of managing intellectual property rights and authentication.

## **U.S. Environmental Protection Agency/ Environmental Information Management System - Robert Shepanek**

### **Purpose**

EIMS is an inter-active system used to support all aspects of the analytical process. The system houses metadata descriptions of a wide variety of information objects including, projects, data sets, databases, models, algorithms, images and multi-media products. Users of EIMS are able to search these descriptions, review the meta-data to determine if the information object has value for their purpose and in many cases, then retrieve and use the object.

Although designed to interactively support the analytical process, EIMS with some modification will support the long-term archiving of data and other information objects. A project is currently underway within EPA to integrate EIMS into the agency-wide records management approach.

### **What is Archived?**

EIMS focuses on data, information products, and tools that are useful in environmental health and ecological assessment. This is a broad definition that may include information from a variety of disciplines such as economics, transportation and others.

EIMS is an Oracle database that contains the metadata descriptions for all of the objects tracked and made available by the system. In its current configuration, data sets selected for use in an EIMS supported project are actually loaded into the analytical component of the Oracle database that also houses the metadata component. The purpose of copying data into the common structure, for example merging data sets of different types like SAS or spreadsheets, is to facilitate analysis of it by multiple investigators leveraging a common structure and data type. Data, information and tools that are not in the database are made available by FTP. In some cases, the links may be to data, information and tools at external sites.

EIMS is predicated on a partnership model. Responsibility for data administration and stewardship is distributed to the organization that owns the object. Refresh and archival of external objects that are linked to metadata descriptions in EIMS are subject to the policies and procedures in place within the partner organization.

### **Format**

Metadata is captured in Oracle native format using a Web forms front-end. As reflected previously in the narrative, data are in multiple formats including Oracle. Given the objectives of the system, transformation of the data for very-long-term storage are under investigation. These considerations are less important for those data that are actually in use and in the EIMS database component as migration paths exist for the Oracle database. Transformation of data referenced in the system but stored externally is a critical issue that is under investigation.

### **Metadata**

EIMS has a robust metadata management component that has been developed to meet the, “20 year rule”.

“Will someone 20 years from now, not familiar with the data or how they were obtained, be able to find data sets of interest and then fully understand and use the data solely with the aid of the documentation archived with the data set?”

The system is on the brink of full compliance with the Federal geographic Data Committee

metadata standard. Metadata in EIMS are organized at 3 levels.

**S**     Directory level

Assist the user in identifying data sets of interest and making an initial determination of whether they are useful for his intended purpose.

**S**     Catalog level

Assist the user in performing an in-depth evaluation of a data set or tool.

**S**     Dictionary level

Provide detailed information about individual attributes.

**S**     Data level,

Based on project specific needs, for example multiple data sets required for the analysis by multiple investigators, data sets are migrated into the common database structure facilitating analysis without losing the full data pedigree, and connection to the metadata housed in the levels referenced above.

**Process**

EIMS is based on a partnership model. Creators, who remain the owners and stewards of the data and metadata, currently include a variety of programs within EPA, EPA Regions and one state. In some cases EIMS points to data held by other federal and non-federal organizations. A data administrator manual is currently under development for EIMS that will provide the policies, procedures and standards for partners dealing with data administration issues including archival.

**Technology**

Data made available through EIMS come in a variety of software and on multiple types of media. For example, data are described in EIMS that are stored on CD-ROM. These data are obtained by sending a request to the owner for a copy of the data. Data and metadata are also housed in the Oracle database. It is expected that as EIMS moves to more advanced versions of Oracle that more data types will actually be stored in the database. For example, Oracle 8 will support storage and indexing of spatial data in the database. This should simplify the current archival problem

**Costs**

EIMS has been a multi-year effort that began as the systems development effort for the EPA Environmental Monitoring and Assessment Program (EMAP). The user base for the system has greatly expanded since that time, and the system has continued to evolve based on new user requirements. Current year expenditures for system enhancement are running in the 400K range. Operational and database administration costs are roughly 100K per year. Data administration and content development costs are borne by the EIMS partners and vary greatly depending on how active the partner is in content development.

## **Policies**

Data administration policies, standards and procedures are under development. In order to be an EIMS partner, organizations must provide a data librarian whose focus is on the development and quality of content. There is a set mandatory data elements that must be provide in order to enter a metadata record into EIMS. Within the scope of the data administration policies and procedures, individual partners have considerable latitude in how they manage their data administration program.

## **Access**

EIMS metadata, data, information and tools are accessible using web based technology. A web browser is all that is required to review the metadata and to download data and tools. The current direction of EIMS is to add suites of statistical, spatial and visualization routines that can be used to access both data and metadata. Many of these routines will be invoked through knowledge based front-ends. Long-term archive accessibility is currently under design and will be integrated with the EPA records management approach.

## **Future Plans**

EIMS will become a comprehensive inventory of a wide variety of environmentally related information and tools as the list of engaged partners grows. In addition, it is slated to become the core of the ORD Science Information Management System. This system will include many of the features already addressed in this narrative, as well as greatly enhanced analytical capability, a robust very long-term archival capability and interoperability with relevant systems inside and outside of EPA.