e r p a e n

Information Society Technologies

netherlands historical data archive

ELECTRONIC RESOURCE PRESERVATION AND ACCESS NETWORK

erpanet

www.erpanet.org

ERPANET – Electronic Resource Preservation and Access Network – is an activity funded by the European Commission under its IST programme (IST-2001-3.1.2). The Swiss Federal Government provides additional funding.

Further information on ERPANET and access to its other products is available at http://www.erpanet.org.

A great deal of additional information on the European Union is available on the Internet. It can be accessed through the Europa server (http://europa.eu.int).

ISSN 1741-8682

ELECTRONIC RESOURCE PRESERVATION AND ACCESS NETWORK

Table of Contents

Executive Summary

The profession of Data Archiving has a business driven objective in ensuring information can be successfully preserved and was one of the first professions to encounter and face the challenges of preserving digital information. This case study on the Netherlands Historical Data Archive (NHDA) in Amsterdam, which has been collecting and preserving datasets relating to Dutch historical research for almost fifteen years, serves as an example of the work being carried out in this profession. The NHDA has a growing collection of over 650 data collections, of varied size and complexity, plus a smaller collection of images and websites. This case study focuses on the NHDA's activities in preserving their datasets, the mainstay of their collection.

The NHDA are active in monitoring new developments and in keeping pace with professional and international standards and practices. Their strategy is thus broadly consistent with other national data archives and has been tailored to their situation as a historical and partially bi-lingual archive. They have devised a strategy and procedure for converting datasets into a standard format that can be read back into contemporary applications when required by users. The most frequently requested datasets are copied onto the Institute webserver and provided online for free consultation, provided that the depositor has given permission.

The strategy at the NHDA is informed not only by other professional bodies in the sector but also by their experiences in pilot projects carried out to research potential new strategies or practices. Particularly notable are their collaborative efforts in developing a Dutch derivative of the Data Documentation Initiative (DDI), the DDDI, and the recent eXtensible Past (XPast) project, which examines opportunities for converting their archival holdings and metadata to XML. The NHDA consider themselves to be at the forefront of digital preservation research for data archives and are keen to maintain a proactive role engaging in new and innovative projects where possible and relevant.

Chapter 1: The ERPANET Project

The European Commission and Swiss Confederation funded ERPANET Project[1] (Electronic Resource Preservation and Access Network) works to enhance the preservation of cultural and scientific digital objects through raising awareness, providing access to experience, sharing policies and strategies, and improving practices. To achieve these goals ERPANET is building an active community of members and actors, bringing together memory organisations (museums, libraries and archives), ICT and software industry, research institutions, government organisations, entertainment and creative industries, and commercial sectors. ERPANET constructs authoritative information resources on state-of-the-art developments in digital preservation, promotes training, and provides advice and tools.

ERPANET consists of four partners and is directed by a management committee, namely Seamus Ross (HATII, University of Glasgow; principal director), Niklaus Bütikofer (Schweizerisches Bundesarchiv), Hans Hofman (Nationaal Archief/National Archives of the Netherlands), and Maria Guercio (ISTBAL, University of Urbino). At each of these nodes a content editor supports their work, and Peter McKinney serves as a co-coordinator to the project. An Advisory Committee with experts from various organisations, institutions, and companies from all over Europe give advice and support to ERPANET.

[1] ERPANET is a European Commission funded project (IST-2001-32706). See www.erpanet.org for more details and available products.

Chapter 2: Scope of the Case Studies

While theoretical discussions on best practice call for urgent action to ensure the survival of digital information, it is organisations and institutions that are leading the drive to establish effective digital preservation strategies. In order to understand the processes these organisations are undertaking, ERPANET is conducting a series of case studies in the area of digital preservation. In total, sixty case studies, each of varying size, will investigate awareness, strategies, and technologies used in an array of organisations. The resulting corpus should make a substantial contribution to our knowledge of practice in digital preservation, and form the foundation for theory building and the development of methodological tools. The value of these case studies will come not only from the breadth of companies and institutions included, but also through the depth at which they will explore the issues.

ERPANET is deliberately and systematically approaching disparate companies and institutions from industry and business to facilitate discussion in areas that have traditionally been unconnected. With these case studies ERPANET will broaden the scope and understanding of digital preservation through research and discussion. The case studies will be published to improve the approaches and solutions being developed and to reduce the redundancy of effort. The interviews are identifying current practice not only in-depth within specific sectors, but also cross-sectorally: what can the publishing sector learn from the aeronautical sector? Eventually we aim to use this comparative data to produce intra-sectoral overviews.

This cross-sectoral fertilisation is a main focus of ERPANET as laid out in its Digital Preservation Charter.[2] It is of primary importance that disparate groups are given a mechanism through which to come together as best practices for digital preservation are established in each sector.

Aims

The principal aims of the study are to:

- build a picture of methods and match against context to produce best practices;
- accumulate and make accessible information about practices;
- identify issues for further research;
- enable cross-sectoral practice comparisons;
- enable the development of assessment tools;
- create material for training seminars and workshops; and,
- develop contacts.

Potential sectors have been selected to represent a wide scope of information production and digital preservation activity. Each sector may present a unique perspective on digital preservation. Organisational and sectoral requirements, awareness of digital preservation, resources available, and the nature of the digital

[2] The Charter is ERPANET's statement on the principles of digital preservation. It has been drafted in order to achieve a concerted and co-ordinated effort in the area of digital preservation by all organisations and individuals that have an interest and share these concerns.
http://www.erpanet.org/www/content/documents/Digitalpreservationcharterv4_1.pdf

object created place unique and specific demands on organisations. Each of the case studies is being balanced to ensure a range of institutional types, sizes, and locations.

The main areas of investigation included:

- perception and awareness of risk associated with information loss;
- understanding how digital preservation affects the organisation;
- identifying what actions have been taken to prevent data loss;
- the process of monitoring actions; and,
- mechanisms for determining future requirements.

Within each section, the questions were designed to bring organisational perceptions and practices into focus. Questions were aimed at understanding impressions held on digital preservation and the impact that it has had on the respective organisation, exploring the awareness in the sector of the issues and the importance that it was accorded, and how it affected organisational thinking. The participants were asked to describe, what in their views, were the main problems associated with digital preservation and what value information actually had in the sector. Through this the reasons for preserving information as well as the risks associated with not preserving it became clear.

The core of the questionnaire focused on the actions taken at corporate level and sectoral levels in order to uncover policies, strategies, and standards currently employed to tackle digital preservation concerns, including selection, preservation techniques, storage, access, and costs. Questions allowed participants to explore the future commitment from their organisation and sector to digital preservation activities, and where possible to relate their existing or planned activities to those being conducted in other organisations with which they might be familiar.

Three people within each organisation are targeted for each study. In reality this proved to be problematic. Even when organisations are identified and interviews timetabled, targets often withdrew just before we began the interview process. Some withdrew after seeing the data collection instrument, due in part to the time/effort involved, and others (we suspect) dropped out because they realised that the expertise was not available within their organisation to answer the questions. The perception of risks that might arise through contributing to these studies worried some organisations, particularly those from sectors where competitive advantage is imperative, or liability and litigation issues especially worrying. Non-disclosure agreements that stipulated that we would neither name an organisation nor disclose any information that would enable readers to identify them were used to reduce risks associated with contributing to this study. In some cases the risk was still deemed too great and organisations withdrew.

Chapter 3: Method of Working

Initial desk-based sectoral analysis provides ERPANET researchers with essential background knowledge. They then conduct the primary research by interview. In developing the interview instrument, the project directors and editors reviewed other projects that had used interviews to accumulate evidence on issues related to digital preservation. Among these the methodologies used in the Pittsburgh Project and InterPARES I for target selection and data collection were given special attention. The Pittsburgh approach was considered too narrow a focus and provided insufficient breadth to enable full sectoral comparisons. On the other hand, the InterPARES I data collection methodology proved much too detailed and lengthy, which we felt might become an obstacle at the point of interpretation of the data. Moreover, it focused closely on recordkeeping systems within organisations.

The ERPANET interview instrument takes account of the strengths and weaknesses from both, developing a more focused questionnaire designed to be targeted at a range of strategic points in the organisations under examination. The instrument[3] was created to explore three main areas of enquiry within an organisation: awareness of digital preservation and the issues surrounding it; digital preservation strategies (both in planning and in practice); and future requirements within the organisation for this field. Within these three themes, distinct layers of questions elicit a detailed discovery of the state of the entire digital preservation process within participants' institutions. Drawing on the experience that the partners of ERPANET have in this method of research, another important detail has been introduced. Within organisations, three categories of employee were identified for interview: an Information Systems or Technology Manager, Business Manager, and Archivist / Records Manager. In practice, this usually involved two members of staff with knowledge of the organisation's digital preservation activities, and a high level manager who provided an overview of business and organisational issues. This methodology has allowed us to discover the extent of knowledge and practice in organisations, to understand the roles of responsibility and problem ownership, and to appreciate where the drive towards digital preservation is initiated within organisations.

The task of selecting the sectors for the case studies and of identifying the respective companies to be studied is incumbent upon the management board. They compiled a first list of sectors at the very beginning of the project. But sector and company selection is an ongoing process, and the list is regularly updated and complemented. The Directors are assisted in this task by an advisory committee.[4]

[3] See www.erpanet.org. We have posted the questionnaire to encourage comment and in the hope that other groups conducting similar research can use the ideas contained within it to foster comparability between different studies.
[4] See www.erpanet.org for the composition of this committee.

Chapter 4: Introduction to the data archiving sector

The NHDA is the main Dutch organisation for archiving historical data. A few other organisations in the Netherlands are also concerned with data archiving, such as the Steinmetz Social Sciences Data Archive, but most of the NHDA's contemporaries are found overseas. Most Data Archives hold datasets created and contributed by academic researchers in a specific field. The datasets are contributed so that they can be re-used by other researchers in new studies and comparative research, and together they form a valuable record of specific fields of research.

Data archiving as a sector is largely unregulated. An associate member of the International Social Science Council of UNESCO, the International Federation of Data Organisations (IFDO), offers guidelines and a selection of papers on setting up and maintaining a digital archive, but it is not the role of this organisation to specify legal requirements and so the guidance remains simply as guidance. The other major international data archiving organisation is CESSDA, the Council for European Social Science Data Archives. As much of the social science archiving infrastructure can be borrowed and adapted to fit the needs of the historical archives, data archives and social science archives tend to maintain strong links. In this respect, the International Association for Social Science Information Service and Technology (IASSIST) is also influential in promoting a network of excellence for data service delivery and providing opportunities for sharing and transferring knowledge across different institutions.

The sector of data archiving, specifically social science and historical data, is generally more advanced in the active implementation of digital preservation than many other sectors. This can be put down largely to a) the nature of the sector having archiving as its mandate, and b) the relatively long-standing presence of such data archives (at least insofar as computing history is concerned), some of which have been archiving manipulable digital data since the 1960's.

Data archives have a business driven objective to preserve their material. Levels of awareness and understanding are generally high, as is practical experience. It is perhaps therefore understandable that our interviewees perceived the main challenge not so much in the rapid change of pace in technological developments that so quickly make formats and media obsolete, but the amount of labour and man hours required to keep up with it.

Chapter 5: Details of the Interviews

The Netherlands Historical Data Archive (NHDA) is maintained by the Department of History at the Netherlands Institute for Scientific Data Services (NIWI), established in 1997 and based at Amsterdam[5]. NIWI is an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW).

The NHDA

http://www.niwi.knaw.nl/en/geschiedenis/collecties/toon

The initiative to set up a historical data archive for historical research in the Netherlands began as a project at Leiden University in 1989. In 1995 the Archive was incorporated into the KNAW and in 1997 it merged with five other Academy institutes, including the Steinmetz Social Science Data Archive (formerly a unit of the SWIDOC[6]), to form the NIWI at KNAW. Its close proximity to these other institutes provides valuable ground for exchange of knowledge and learning.

The NHDA collects historical datasets, CD-ROMs and websites with relevance to Dutch history and makes them available for further research[7]. They have some 20,000 images stored on CD-ROM and online, and a small number of websites (4 available online in October 2003) of re-presented historical datasets with context and related materials. Their main concern however, rests with the historical datasets themselves. In September 2003, their catalogue of historical datasets contained 656 data collections clustered in seven subjects[8]. The size of the datasets ranges from very small (i.e., datasets from individual dissertations with just a few hundred rows/records) to very large (up to 100,000 rows/records in some collections).

Datasets are deposited willingly and the Archive does not follow an aggressive policy of acquisition.

[5] NIWI was formally started in 1997, the result of the merger of six previously existing institutes providing scientific information in the fields of social sciences, history, and Dutch language and literature. In addition to the NHDA, the NIWI incorporates another data archive -the Steinmetz Social Sciences data archive -and maintain online databases on Dutch Research and Social Research Methodology.

[6] Social Science Information and Documentation Centre

[7] The information is often made freely available online or can be consulted in a personal visit to the premises.

[8] Datasets are categorised as follows: economic history; social history; demographic history; prosopography; history of trade and navigation; history of culture; and others.

Chapter 6: Circumstances

The NHDA maintains a relatively high level of activity in the cutting edge of digital preservation despite their small numbers. They have a small but dedicated staff, many of whom are also involved in digital preservation projects inside and outside of their NHDA work. Most of the team are specialists in data archiving.

ERPANET initially contacted Annelies van Nispen, data archives specialist and project manager of the NHDA's XPast project (described below). The NHDA were keen to be involved and provided the time and services of Peter Doorn, head of the History Department and Director of the NHDA, and Marjan Balkestein, senior data archiving specialist.

The interviews took place at the NHDA at NIWI, Amsterdam, at the end of September 2003.

Chapter 7: Analysis

This section presents an analysis of the data collected during the case study. It is organised to mirror the sequence of topics in the questionnaire.

- Perception and Awareness of Digital Preservation
- Preservation Activity
- Compliance Monitoring
- Digital Preservation Costs
- Future Outlook

Perception and Awareness of Digital Preservation

Digital preservation is recognised as a significant issue by the NHDA, stemming from their origins as a digital data archive. They monitor a wide variety of sources for new digital preservation developments and are keen to participate in relevant activities, for example the DLM Forum[9] and ERPANET workshops[10]. They have good knowledge and overall awareness of the main facets of digital preservation (such as metadata, technical strategies, and standards) and have implemented procedures and strategies across the Archive for maintaining their digital data through time.

Over the last few years, the History department has carried out a number of projects on digital archiving and the digitisation of paper-based material. Having concluded collaborative work on developing the Dutch Data Documentation Initiative (DDDI), they are currently working on the 'eXtensible Past' (XPast) project[11]. They are also collaborating with the NIWI Meertens Institute[12] on the project 'Archiving Digital Academic Heritage' (ADA)[13].

The Main Problems

The NHDA's experience over the past 15 years has given them a broad understanding and appreciation of the different facets of preserving digital information. Like many, they perceive the central problem of digital preservation to be the general process of technological innovation that quickly makes formats and media obsolete. Given that the NHDA have developed extensive experience in a range of data formats and standards over the years, for them this becomes not so much of a technical challenge but a human one. The work needed to maintain familiarity and pace with these changes is highly labour intensive and requires a large number of man hours per year to document, archive, research, and keep their procedures up-to-date.

[9] DLM Forum 2002: http://europa.eu.int/historical_archives/dlm_forum/index_en.htm. The DLM Forum is held once every three years and stands for Document Lifecycle Management. This name was attributed to DLM in 2002; previously the acronym stood for the French 'Données Lisibles par Machine/Machine Readable Data'.
[10] See http://www.erpanet.org
[11] See http://www.niwi.knaw.nl/nl/geschiedenis/projecten/xpast_copy1/toon
[12] See http://www.meertens.knaw.nl/indexe.html
[13] See http://www.niwi.knaw.nl/en/geschiedenis/projecten/ada/toon/

Metadata and access are two issues of increasing importance to the NHDA and research efforts have been established to address them. Metadata was the subject of the 1996 Dutch Data Documentation Initiative (DDDI) project, in which the Steinmetz and Historical data archives collaborated on an interoperable Dutch derivative of the DDI. The DDDI schema incorporates all required metadata fields and is a comprehensive listing of dataset metadata. Work is currently taking place on an XML implementation of the DDDI so that variable dataset metadata can be preserved in direct association with the dataset files, not possible within the Archives' current preservation strategy. Regarding access, which is currently limited for the majority of datasets to the provision of ASCII files that require rendering in the researchers own database application or spreadsheet programme[14], improving access to datasets is one of the research areas of the current XPast project. The aims of the XPast project are twofold: firstly to explore the possibilities of XML[15] and the Open Archives Initiative (OAI)[16] for providing better access to and sharing of digital data collections by researchers; and secondly to investigate XML as part of a new strategy for the long-term preservation of research data. The results of the project are therefore also applicable to their technical strategy for preserving data.

In addition to metadata and access, acquisition is viewed as a potential problem area as the Archives do not pursue an active acquisition policy and rely instead on researchers to freely deposit their data. They are aware of the repercussions of such an approach and the potential gaps in their holdings, and are actively discussing alternatives.

Asset Value and Risk Exposure

The NHDA are aware of the risks they run should they fail to preserve their archival data. Preservation of the data is part of their central mandate as an Archive. However, there is no legal framework for depositing or preserving historical research data in the Netherlands. The risks they face are therefore not legal but business-driven. Their core business would be adversely affected if the information was not preserved properly and their funding and existence would be under threat. From a historical perspective, they would also endanger a valuable record of Dutch research history.

The department is keen to see its data re-used and attributes a great deal of value to its assets in the Archive but they do not carry out official risk analysis or business needs analysis studies. Instead they carry out projects and pilots that incorporate these aspects into their scope. They recently carried such out a pilot project on archiving the records of the Meertens Institute[17]. Such projects are both procedural and actual, providing the Department with good experience that can be re-used in real time full-scale projects.

The department affords higher priority to the maintenance and preservation of its archival holdings than its own administrative data. Strategies for preserving administrative data have not been developed and were thus not addressed again during the interview.

[14] A small number of datasets are available as websites; this is not, however, the NHDA's main approach to making datasets available and have been created as one-off projects by individual researchers/trainees.
[15] XML stands for extensible Mark-up language. See http://www.w3.org/XML/.
[16] The Open Archives Initiative: http://www.openarchives.org/.
[17] This pilot project was part of the larger project, ADA (op cit). This report is due to be published (in Dutch) in 2004.

Preservation Activity

Staff at the NHDA are active in digital preservation. They monitor new developments through a variety of sources such as conferences, workshops, listservs and e-journals, contributing where relevant, and they maintain a high level of awareness about national and international developments in digital preservation. Some of the staff are involved in external preservation projects such as on the preservation of digital images, and others contribute directly to organisations such as the DLM Steering committee. They are keen to learn from the work of others and see research into other existing projects as a vital part of carrying out their own research projects.

Collaboration and communication with related institutions is highly valued. The NHDA draws on the experience of national archives, data archives, and the library community in developing its approach.

Policies and Strategies

Despite the fact that the NHDA has a business-driven objective in dealing with preservation, they do not have an explicit or separate digital preservation policy. The preservation aspect is implicit in general organisational policies, and staff and directors are happy with that. There was an attempt some years ago to write their strategies into a handbook, but it was a significant investment in time for the benefit of a small number of people and some of it had already been outdated before it was finished. This practice is thus no longer maintained.

Their overall approach and strategy is kept up to date by thoroughly monitoring recent developments in digital preservation and tracking the work of their colleagues in other data archives (such as the Steinmetz Archives). The NHDA have a good grasp on digital preservation issues and do not have a specific time frame for reviewing policies and strategies but simply use their own informed judgement to address it as and when required[18]. Keeping an eye on new technological developments ensures that they do not encounter obsolescence problems. They are keen to explore new preservation approaches but will not change their strategy without good reason and thorough preparation. For example, the outcome of the XPast project may well form the backbone of their new policy, depending on the benefits the project can deliver.

Selection

Datasets must meet a basic set of acquisition or selection criteria to be accepted by the NHDA. Despite this, the NHDA do not pursue an active policy of acquisition and datasets are simply contributed directly to the NHDA by willing parties ad hoc.

Selection criteria are not expressed within an explicit selection- or acquisition-policy but have been included in the NHDA annual plans and are communicated to users via the website and other publicity material[19]. The central criteria are that the dataset must be an original one and must have been created in a scientifically acceptable and reliable way. These criteria being met, potential depositors can then submit their research as long as it meets the criteria of *either* being researched by Dutch researchers *or* is on the subject of Dutch or former colonial interests. This restriction

[18] This may prove difficult to maintain as the archive grows and responsibilities become diversified across a larger team.
[19] See for example the NHDA website: http://wwwoud.niwi.knaw.nl/us/dd_nhda/nhdadepo.htm.

has been agreed in conjunction with similar organisation abroad. Lastly, the NHDA will only archive files accompanied by substantive content and technical documentation or metadata. The depositor is responsible for providing such information before the data will be finally accepted.

Transfer into the NHDA follows a tightly defined procedure. Depositors can download the deposit forms from the NHDA web pages and must supply metadata, or catalogue information for each datafile in the set, as well as an overview of the dataset contents[20]. Metadata accords to the Dutch Data Documentation Initiative (DDDI) schema developed by the NHDA in conjunction with the Steinmetz Archives in the late 1990's.

The DDDI is a metadata standard derived from the Data Documentation Initiative (DDI) [21], an international metadata standard for describing social science datasets. The DDDI has been adapted for use in both historical *and* social science archives in the Netherlands. DDDI metadata describes and regulates the content, presentation, transfer and preservation of the datasets at the NHDA. The depositor submits metadata pertaining to content and presentation; administrative metadata concerning transfer and preservation is added by the NHDA. The DDDI is compatible with the international DDI and the Steinmetz Archive is using DDDI-DDI for international metadata exchange. The NHDA are considering the possibilities for doing the same.

Preservation

A dedicated team within the history department is responsible for carrying out preservation activity. They meet regularly to discuss potential problems and are responsible for training new staff members. The Preservation handbook by Beagrie and Jones has proved comprehensive in giving new staff an overview of the main issues associated with preserving digital information[22]. There is very little formal training in preservation processes, although all new members are trained in-house in the procedures and methods the NHDA have established.

The preservation process is centred around a strategy of migration to standard formats, in this case ASCII[23]. The aim of this is to make the datasets as independent as possible from their original hardware/software environment. The datasets are converted from their (often proprietary) format into a standard ASCII fixed format that can later be represented in a basic table. Once the data has been converted to ASCII there is no need for further conversion of the data at the archives. The conversion is documented and the results checked with the original depositor. Images, if present,

[20] Three forms are relevant here: the 'Overdrachtsformulier'; the Inlichtingenformulier Dataset; and the Inlichtingenformulier Data-bestand. The majority of the datasets they receive are from structured databases.
[21] DDI: http://www.icpsr.umich.edu/DDI/.
[22] Jones, M & Beagrie, N, *Preservation Management of Digital Materials* (London, 2001). Available online at http://www.dpconline.org/graphics/handbook/reviews.html.
[23] This is one of the occasions where the requirements of two different institutions play a role in the choice of preservation format, as identified earlier in this section. ASCII is perfectly suited towards the archiving of historical data. The Steinmetz archives, on the other hand, archive their data in SPSS format, a statistical package that operates as a vital portable tool for the social sciences. For further information on the ASCII table, see http://www.w3.org/People/howcome/p/telektronikk-4-93/ascii.html.

are converted into TIFF Group 4 [24](although JPEG[25] has been used on occasions where images have been altered prior to deposit).

The Archives retains the original files after conversion but without any preserved way to access them, and they do not carry out further integrity checks[26]. A small study was launched some years ago to examine the use of checksums but was halted after it failed to highlight any discernible benefits. Users of the datasets are expected to identify any errors and report them. On no occasion has a conversion error been reported.

DDDI Metadata is manually entered by NHDA staff and processing priority is afforded to the datasets from large projects. The department has experimented with automated collection but found it largely incompatible within their current ingest and access process (except for images, where the headers of the TIFF files can be parsed). One of the main stumbling blocks here is language – much of the data is submitted in Dutch but is accessed in English.

The DDDI metadata is the complete set of metadata required to maintain and preserve the dataset files over time. Documentation on the variables and codebooks is not kept in the DDI-system but stored together with other additional data documentation. This may be digital, but sometimes it is only available in paper form. In conjunction with the XPast project, an XML version of the DDDI schema is currently under development that will allow the staff to preserve the metadata directly with the dataset files. This can have multiple benefits, including compatibility between dataset and metadata storage formats, and data storage in one unique environment (instead of distributed).

Physically, the files are stored using a RAID system (Redundant Array of Inexpensive/Independent Disks)[27]. Shadows are put onto active hard disks, which are backed up on tape with new accessions and which are then copied onto CD-ROM. The CD ROMs are for backup only, and are not intended for active use. A second copy on CD-ROM is also kept offsite in a different location, and new replacement CD-ROMs are made each year. Refreshed CD-ROMs are made from files on the active hard disks, not the CD-ROM backup copies. Maintenance of the hardware and software is not directly attributed to the NHDA staff but is the responsibility of the NIWI IT department. Knowledge of digital preservation issues amongst this group is not as advanced as that of the NHDA team.

The staff at the NHDA do not consider digital storage to be a challenging issue, as storage is getting cheaper as time passes and more data is accrued. Basic storage is no problem for them as long as the media are regularly refreshed[28]. The issue of media refreshment has caused them more problems than bit deterioration, largely

[24] TIFF stands for Tagged Image File Format and is a common, flexible raster file format. Although owned by Adobe, the specification is published and online at http://partners.adobe.com/asn/developer/pdfs/tn/TIFF6.pdf.
[25] JPEG stands for Joint Picture Expert Group (so-named after the committee that originally wrote it) and is a standardized image compression system mechanism. See http://www.jpeg.org/.
[26] Some users prefer to receive the data in its original format, which the NHDA caters for in preserving these files. The computer museum at Amsterdam has been used to gain access to these files, on more than one occasion. The staff also feel that these files may have future uses as yet unknown.
[27] See: http://redundantarrayofindependentdisks.com/.
[28] They have encountered only one possible problem with bit deterioration in a dataset from the 1960 census, which may have been present even when the dataset was transferred.

because they refresh the data over their network and their backup system is not set-up to handle large amounts of data in that way. They readily suggested that this relaxed attitude to storage may be due to the size of their holdings, being relatively small compared to those of some National Archives. Ultimately however, they consider the issues of organisation, documentation and accessibility to be far more important than storage.

The team has selected their approach on the basis of research, practical trials, and evaluation. The standards and basic technical strategies they employ are compatible with the basic practices followed at many other major data archives.

Access

To provide public access to the datasets, copies of the files are made from the ative hard disk and are transferred onto the NHDA webserver for access over the Internet. Dublin Core access metadata will be incorporated into the DDDI schema. The catalogue is fully browsable by time period, geographic area, discipline, keywords, researcher (depositor/author), source, and publication.

Approximately a third of the data is permanently available online from the webserver. The remainder can be obtained by submitting a request to staff at the NHDA, although they rarely receive requests for datasets that are not already available online[29]. The datasets are supplied in the format they have been preserved in – flat ASCII files - and no application facilities are provided. Users must import the data into a suitable application themselves.

The data held by the NHDA is prevented from unauthorised access and manipulation as part of the data protection levied by the IT support staff across the NIWI. Access to critical areas is restricted to members of the NHDA. There have been no known instances of information altering.

Some copyright issues have been faced with recent digitisation projects, although the NHDA do not face significant copyright issues as most of their source material is public or free from copyright. Privacy issues are dealt with in co-operation with the Dutch Data Protection Authority, as is the case with datasets referring to people[30]. In a project to archive and make available the population censuses of 1960 and 1971, access restrictions were compiled by the responsible authority 'Statistics Netherlands'. Access security and privileges are set when the data is accepted into the archives, with additional role based privileges for staff members.

Access provisions are under review, specifically within the context of the XPast project.

Compliance Monitoring

The preservation process is evaluated as an integrated part of the overall services provided by the NHDA and not as an individual procedure. NIWI's Scientific Committee evaluate and advise on aspects of the service as and when required. In addition to this, every academy institute undergoes a regular audit every five years.

[29] Indeed, they estimated up to 95% of data used is that which is freely available on the web. This may be due to the growing popularity of hobby history, for example, or a lack of patience on the behalf of the users to wait for alternative files and sources.

[30] This is often done in conjunction with other authorities as well; for example, 'Statistics Netherlands', was responsible for defining access restrictions in a project to archive and make available the censuses of 1960 and 1971

This is effectively a peer review that examines all aspects of the Institute, examining strategy and output but leaving process in internal hands. The Academy requires this evaluation from all of its Institutes as a basic Business Requirement.

Materials are not checked for integrity on a regular basis: basic visual checking takes place only before materials are issued and user de-facto evaluations are considered to take place each time a user uses the material. Checking each and every dataset is deemed highly inefficient unless the process can be automated, which the NHDA have so far been unable to do. The Archives thus relies heavily on users to inform them of any problems with the data.

Digital Preservation Costs

Preservation is not seen as a separate cost activity by the NHDA, or the NIWI. The costs of preserving data are simply part of the overall NHDA portion of the budget, all of which is dedicated to maintaining the NHD archive in some way. Most archiving carried out by the NHDA is paid for directly by NIWI. Some further funds can be obtained for specific projects from third party sources such as the SURF foundation[31].

The NHDA do not undertake cost benefit analyses for their new projects; costs are instead factored into pilot studies. Both the ADA and the XPast projects are incorporating cost factors as a study element.

Future Outlook

The future of the Netherlands Historical Data Archives is uncertain, at least in its current form as a NIWI institute. It was recently announced that the Royal Academy intends to establish a national data archive as an independent institute, in which at least the NHDA and Steinmetz Archives will be incorporated. The new structure is expected to become operational in 2005.

At the time of interview (before the announcement of institutional restructuring, above), the NHDA were optimistic about the future. They predicted their current preservation activities would be sufficient as long as they continued to meet the needs of their users. If the needs of the users were met then, by and large, so were those of the organisation. They considered their archiving process to be very limited in terms of restructuring potential, as it would take an inordinate and near heroic amount of effort given the scale of the work involved and the limitations on resource availability.

The NHDA see data archiving not a passive activity but one that can, and should, be responded to as circumstances change and as new developments or opportunities present themselves. It is in line with this pro-active attitude that the team have established and collaborated on projects researching new approaches to digital preservation and data archiving procedures. They believe that they could actually continue securely using ASCII for several more years to come but that it is no longer the format of choice for users. User demands are very important, and the NHDA consider it important to base their approach on one that is compatible with user needs. XML is therefore being investigated for multiple purposes with the overall aim of improving their holdings and access to them.

[31] SURF is the higher education and research partnership organisation for network services and information and communications technology: http://www.surf.nl/home/. The NHDA are eligible for funds as they are working for the benefit of higher education and research.

The NHDA are keen to see investment in digital archiving and preservation increase in the near future. The staff at the Archives are becoming increasingly diverse in their job skills and the director sees their role as not simply that of a stagnant data archive, but to engage in innovative projects. Resources, both financial and human, are needed to ensure they continue as the department hopes. Not only are staff working on the XPast and ADA projects, but they are also considering studying developments in Archaeological Archiving (as currently on trial in New Zealand) to see if they can carry out a trial for the Archaeological profession in the Netherlands, amongst others[32]. They are aware that they can learn from other individuals and organisations and are keen and willing to do so where possible.

[32] See http://www.niwi.knaw.nl/en/geschiedenis/projecten/toon/ for further NHDA projects

Chapter 8: Conclusions

Preserving digital information is a task well managed within the NHDA, which has a business driven objective in ensuring the data is securely and effectively preserved. The Archive was established to deal with digital material and the challenges of digital preservation are ingrained into the Archives overall approach and strategy.

Both staff and management are keen to research new issues and develop ways to secure the content of their archive through time. Their strategy of testing aspects of preservation with pilot projects and studies such as XPast and ADA provides them with valuable practical experience that contributes to future development and improvement of the Archive. Smaller projects, such as those to create dataset websites to address the lack of application facilities provided for the datasets whilst providing a very straightforward method for browsing the datasets online, provide closer insight into different preservation options and complement their larger research efforts. It also exemplifies their desire to research and try out innovative and alternative ways of preserving and making accessible their data.

Exchanging experiences and practices with other data archives, national archives and the library community, results in a high level of interoperability and compatibility with other data organisations that maintain similar holdings. The NHDA are part of a distributed network, insofar as they have adapted or adopted industry-wide standards for metadata and dataset formats. Their research into XML for use in archival holdings is complimentary to other, similar research efforts[33] and reflects the general direction of research into preserving digital information at an international level. The NHDA publish material about their preservation activities in Dutch *and* English, allowing the English speaking communities to benefit from their research as well as Dutch speakers

Their overall preservation strategy is sound; although some may dispute the effectiveness of an evaluation procedure in which users are relied upon to report any errors to the Archive. As the scale of the NHDA holdings can only be expected to increase in the future, the NHDA may be able to use the opportunity of the pilot projects described in this paper to research development of an alternative, automated procedure. They could also apply some of the basic lessons they have learnt in preserving their archive to their administration. Whilst their administration comprises very different data types compared to their archival holdings, many of the basic principles of selection, appraisal, storage and metadata still apply. The attention afforded to this area is comparable to that which many businesses afford *their* archive, reflecting the importance that core business activities have when compared to supporting or administrative ones.

The NHDA‘s datasets are mostly of a static nature and they are not faced with the challenge of preserving active, dynamic databases, as many other organisations

[33] The Dutch Digital Preservation Testbed is the most local example of these, see http://www.digitaleduurzaamheid.nl.

are[34]. However, this should not detract from what they have achieved, namely the successful preservation of a wide variety and number of datasets over time and provision of a research platform into new ways of digitally preserving and accessing dataset files and metadata.

[34] This was the subject of many presentations during the Erpanet workshop on Preserving Databases in Bern, April 2003. See for example, Niklaus Bütikofer's presentation: 'Archiving snapshots or transactions: extracting the right data at the right time from temporal databases' available online at http://www.erpanet.org.

erpanet

CONTACT DETAILS

ERPANET Coordinator

George Service House
11 University Gardens,
University of Glasgow
Glasgow, G12 8QQ,
Scotland

Tel: +44 141 330 4568
Fax: +44 141 330 3788
Coordinator@erpanet.org

ERPANET STAFF

directors
Seamus Ross, Principal Director
Niklaus Bütikofer, Co-Director
Mariella Guercio, Co-Director
Hans Hofman, Co-Director

coordinator
Peter McKinney

editors
Andreas Aschenbrenner
Georg Büchler
Joy Davidson
Samir Musa
Maureen Potter

www.erpanet.org

ELECTRONIC RESOURCE PRESERVATION AND ACCESS NETWORK