*erpa*workshop

The Long-term Preservation of Databases

ERPANET Workshop Report, Bern
April 9 - 11, 2003

Table of Contents

Executive Summary

On 9 to 11 April 2003, ERPANET hosted an expert workshop on the long-term preservation of databases. The workshop took place at the Swiss Federal Archives in Berne and assembled 65 participants from 16 countries all over Europe and the USA. As the third in a series of six expert meetings, the workshop aimed at gathering knowledge about the state-of-the-art in database preservation, fostering discussion and exchange, and at promoting and developing new approaches in the field of digital preservation.

Seven approaches were presented as practical database preservation experience or as development projects with a strong focus on practical solutions. They represented state archives and research projects from across Europe, including an example from the United States. The range was from well-established projects (incorporating many years of experience), to current research in the early stages of implementation, with speakers sharing their experience of developing database preservation strategies. This first part of the workshop intended to provide participants with an overview of the current state of database preservation.

The workshop's second half examined the issue from a different point of view. It presented different steps of the database preservation process, ranging from extraction, appraisal and description to access and adding value. These presentations gave the opportunity to discuss specific aspects of database preservation in greater depth, in particular archival requirements and a scientific data focus on value-adding.

The workshop demonstrated that current development projects will provide archivists with appropriate methods and tools for the long-term preservation of complex relational databases. Presentations and discussions highlighted a series of valuable best practice elements, and also a number of issues that need further research.

Introduction

Databases have been, and continue to be, a key technology for the storage, organisation and interrogation of information. They are a core module in most of today's information systems. While the value of other types of digital information has been highlighted in recent years, databases and the information they contain have often been neglected. Their preservation is of high concern as they are often either irreplaceable or of such value that replacement would be prohibitively expensive.

Databases not only retain the information in a highly structured manner, but they are in most of the cases constantly updated and must allow flexible interrogation of the valuable data. For their long-term preservation, this poses a number of unique problems.

Different communities approach the archiving of databases in very different ways. This ranges from computer science, where archiving usually includes simple backup operations and moving parts of the database into more remote memory, but without a long-term approach, to scientific research databases, where preservation is a necessity, but the focus lies on access, normalisation and value-adding, and to archival science, where long-established techniques and methods as well as the need to maintain authenticity of data meet the technological challenge of databases in order to preserve them for an indefinite period of time.

Aims and Objective of the Workshop

To enable comparison of the various approaches, especially during breakout sessions and conclusions, participants were encouraged to address a range of key questions:

- What underlying preservation policy and/or strategy guides the preservation of databases?
- How is the whole preservation process from extraction to use structured?
- What can be learned from traditional archival appraisal for the selection of databases for preservation?
- What documentation is preserved together with a database, and in what format?
- What users are expected to access the archived databases, and how is this access technically enabled?
- How can archival science and information technology cooperate for efficient database preservation?
- What are the long-term perspectives for the preservation of databases?
- Have any theoretical solutions been completely or partly implemented, and what lessons can these efforts bring to bear on the current state of knowledge?

While speakers and participants represented a focus on state archives and archival science in the main, considerations of computer science, information technology and scientific research all played a role during the workshop. It is evident that the long-term preservation of databases must draw on diverse fields to achieve further insight and experience on this preservation challenge. This workshop's principal goal was to gather

information and experience that is available and foster exchange and communication between different fields of study.

The report is organised into two main sections to mirror the structure of the workshop presentations and discussion. After an introduction to database preservation, the first section presents the practical experience of a range of database preservation projects, while the second focuses on addressing some of the more difficult challenges that have arisen as a result of this experience.

Introduction to Database Preservation Issues

Peter Keller-Marxer, head of the Swiss Federal Archives' ARELDA (Archiving of Electronic Data and Records) programme, delivered the keynote paper on database preservation. The development of databases during the last two decades was outlined, from simple containers of tabulated data to complex systems, now providing full internal data life-cycle management, covering many different data types, and enabling multi-user transaction recording, handling terabytes of data efficiently. Since they have become integral parts of nearly all information systems, databases are also at the technical basis of modern Records Management Systems, a role that provides new challenges for archivists. The value of databases, and thus the importance of their preservation, stems naturally from the importance of the information they contain. This is often irreplaceable or could only be recreated at enormous costs. Increasingly, databases also contain high evidential value, reflecting their supporting role in numerous critical business processes.

There are certain requirements a successful long-term preservation of databases programme has to fulfil. It has to guarantee the integrity, intelligibility, authenticity, originality and accessibility of the preserved databases continuously. These are commonly competitive and conflicting goals in a digital environment in general, and even more so for databases. Key points of consideration are:

Archival Appraisal of databases is only possible through an analysis of the whole information system of which they form part. The purpose, design and context of usage of such a system need to be taken into account. In consequence, reliable appraisal requires extensive technological and systems knowledge. However, this effort is indispensable for sound decisions on what databases need to be preserved. The cost factor of database preservation must also be taken into appraisal considerations.

Extraction of data from databases must distinguish between active and inactive data. Two approaches are common, either to archive snapshots of the entire database, or to archive data marked for deletion. As modern databases are often linked together and form so-called federated databases that can only be archived in isolated parts, an archival decision is needed to define these parts and the way the relations among them are represented. Since the original structure of the data tends to be very complex, referential integrity is a crucial issue. High degrees of complexity are encountered in time-stamped or backlogged databases. In addition, data types and check constraints need to be archived.

Description of databases is often hampered by very poor documentation, which is needed to fully understand extracted data and the context of its creation and use. This is an area where archivists need to raise awareness and to collaborate closely with record-producing agencies. When applying standard description methods which have been developed with only documents and files in mind to databases, significant problems are encountered that call for careful adaptation of common practice.

Preservation methods must aim at extracting the data from their native environment, while still guaranteeing authenticity. Due to the high number of databases expected to be archived in the future, automation of this process will become indispensable.

Access issues are a scale and organisational problem. It is evident that preservation is incomplete without regulated and relatively simple access procedures. While this issue tends to be neglected at the beginning of database preservation projects, access procedures should be included in initial considerations.

Practical Experiences in Database Preservation

The workshop's first session assembled reports from seven database preservation projects. All deal with government data, but offer a wide range of expertise, approaches and perspectives.

Centre des Archives Contemporaines, France

CONSTANCE

The French Programme CONSTANCE (CONServation et Traitements des Archives Nouvelles Constituées par l'Electronique/Preservation and Treatment of New Archives Derived from Computer Processing) has been active since 1979. Located at the Centre des Archives Contemporaines[1] in Fontainebleau, the programme forms part of France's Archives Nationales, the state archives for the central government administration. CONSTANCE archives databases of historical value at the national level. The whole process is set down in detail, and all steps are executed in close collaboration with government agency IT staff. Data are extracted in ASCII character code flat files, and a set of 16 technical metadata elements is created and archived for each file. Extensive documentation comprises data dictionaries, input screens, bibliographical references, manuals and legal documents. By 2002, CONSTANCE held some 6000 files. Different possibilities for access exist, but the respective funding issues have only been partly resolved.

The method used in CONSTANCE represents traditional data archiving which was developed in the 1980s in North America. Even today it is still an appropriate method for archiving simply structured data from surveys etc., and is therefore still used in the scientific and administrative field. Description of the datasets is very much 'hand-made' and therefore very labour intensive.

City Archives of Antwerp, Belgium

DAVID

In Flanders, the DAVID (Digital Archiving in Flemish Institutions and Administrations) research project[2] develops archiving solutions for different kinds of electronic records created by the city administration, aiming at simple, scalable solutions for common environments. These are implemented and tested at the City Archives of Antwerp.

[1] http://www.archivesnationales.culture.gouv.fr/cac/fr/. All Websites visited on 11 June 2003.
[2] http://www.antwerpen.be/DAVID/

The procedure developed by DAVID for archiving records from information systems begins when changes in those systems are made in the administration. The systems can then be appraised, and in cases where archival value is detected, preservation demands are implemented. In preserving databases and information systems, XML plays a major role, as this is considered to be the most appropriate preservation format for structured textual information. However, the translation from a relational data model to a hierarchical document model like XML may be problematic. Geographic Information Systems (GIS) pose some problems at present, but the GML (Geographic Markup Language) standard still under development might offer a solution to the management of this type of system. In general, collaboration with agencies has been smooth and successful, but has so far been limited to simply structured databases like population registers, which are transformed into single XML files.

National Archives & Records Administration (NARA), USA

The Archival Electronic Records Inspection and Control System (AERIC)

The Electronic and Special Media Services Division of the United States National Archives[3] currently holds some 20,0000 files in software-independent form, the bulk of which is survey data. The accessioning process consists of transfer to the archives, first-time copying onto preservation media, verification, and creation of a description package. The AERIC system forms the verification process, which has been operational since 1990, handling basic flat file types. It compares the records' actual content with what is described in the supporting documentation. Through AERIC, validation of the received databases can be handled with maximum efficiency in a semi-automatic fashion.

This verification is crucial for the digital preservation process, since accessioning is usually the last stage for identification and correction of deficiencies and inaccuracies in the records. The many inconsistencies detected in the daily accessioning work at NARA are proof of the importance of such an instrument. For access, databases are handed over to the user in unprocessed form, as the whole archived flat file.

The practice at NARA is very similar to the French method. AERIC is a proven tool for describing and verifying the data structure and the data elements. It therefore facilitates the work of data archivists considerably. However, it supports only one part of the data life cycle and lacks interfaces to access systems, and is not capable of dealing with complex relational databases.

Project Digitale Duurzaamheid, The Netherlands

The Digital Preservation Testbed

The Dutch government's Digital Preservation Testbed project[4] aims at securing sustained accessibility to reliable government information in the digital era. The project has a strong experimental focus, testing several approaches, examining basic requirements, creating and preserving metadata. It focuses on four record types, one of which is databases, and uses three preservation approaches – migration, emulation, and conversion to XML format. Given the existence of different types of databases and the as yet uncertain relationship between databases and records, there are different preservation options available, ranging from preserving single records to the whole database system. The work so far has focused on relational databases and their conversion to XML, with a review of

[3] http://www.archives.gov/
[4] http://www.digitaleduurzaamheid.nl/

commercially available converting tools and development of an alternative. In this approach each database table is converted into one XML file. The relational structure of the whole database can be represented through so-called overview XML files containing information about the structure and the constraints of the database. A working prototype is expected to be ready soon.

University of London Computing Centre, UK

NDAD (National Digital Archive of Datasets)

The University of London Computing Centre operates the National Digital Archive of Datasets under contract to the UK National Archives (formerly PRO)[5] and was established in 1997, becoming the first national archives service to provide online public access to preserved material. Datasets are selected by the National Archives, and remaining activities (acquiring, preserving, describing, providing access and support, and promoting) are handled by the NDAD.

NDAD contains data in the form of datasets as single tables, and contextual material. The data's life-cycle management is modelled on paper records procedures, but certain changes have been made to this. The accession process includes the collection of documentation from records managers and checks for completeness, accuracy and readability. The data are transformed to flat files (while the original is also retained), paper documentation is digitised, and metadata are produced or transformed.

Considerable effort is spent on collecting documentation about the datasets from different sources and checking the consistency of data, while inconsistencies are not corrected, but described. Functionality is documented also. Access is key for the NDAD, and viewing of the datasets is possible on-site as well as via the Internet.

National Archives, Norway

Archiving databases with Arkadukt, Arkade, and ArkN3

The Norwegian National Archives[6] make a distinction between registry-based Electronic Records Management Systems and other systems. The former follow a national standard called Noark,[7] which is in close harmony with international records management standards. It governs the keeping of document-based records and its metadata and therefore archival records can be neatly preserved with a specific tool, called ArkN3. Although the metadata are stored in a database, their standardisation makes it easy to extract them for preservation through a specific interface.

Specialised case-handling systems and other systems each with different data structures require a different approach. The National Archives have developed an Archival Data Description Markup and Manipulation Language (ADDMML), an XML DTD that is used by two tools. The ADDMML-DTD is significant as it can describe datasets consisting of more than one file and table and can therefore deal in a more appropriate way than other approaches with data from complex relational databases. The Arkadukt tool first transfers the original system's description into a hierarchically structured ADDMML file. The Arkade tool then converts, analyses, checks and controls the content of the database based on the ADDMML description file. While doing this it converts the flat preservation files into

[5] http://ndad.ulcc.ac.uk/. The ULCC's Website is at http://www.ulcc.ac.uk/, the National Archives Website at http://www.nationalarchives.gov.uk/.
[6] http://www.riksarkivet.no/english/.
[7] http://www.riksarkivet.no/english/electronic.html.

files for the SAS Software Package, which is in its core module software for statistical analysis but provides not only for verifying the data but also for access and use.

Swiss Federal Archives, Switzerland

Software Invariant Archiving of Relational Databases (SIARD)

The Swiss Federal Archives' SIARD project developed a solution for the preservation of relational databases, building on a strategy that encompasses software independence, standard formats, authenticity and documentation.[8] SIARD defined a preservation workflow and supports it with three software tools. The Analysing and Interpreting tool A0 is used by the record-creating office to check whether the SQL[9] used is compatible with standard SQL-3 compatibility of the database and to extract the archive files using the JDBC interface. The Metadata tool A1 serves to complete the automatically gathered low-level metadata, adding high-level metadata, and will be used by creating agencies and at the archives. Finally, the Reload tool A2 will reload the archive files into a browsable database on a system current at time of access, which, in fact, must be able to import standard SQL-3 and XML files.

SIARD is written in Java to achieve maximum platform independence. It produces three kinds of archive files: a metadata file encoded in XML, an SQL-3 DDL file to represent the structure of the database, and flat files (Unicode 3/UTF-16 encoded) to store the database content. The tools are currently being beta-tested and are expected to be available in Autumn 2003.

SIARD and the method behind this tool represent a considerable step forward as they enable archivists to preserve complex relational databases in an appropriate and highly automated way. The method provides not only for preservation in open and robust formats with good long-term prospects for migration, but also for an easy reload into relational database management systems, thus enabling use and access through many software tools.

Discussion

Some of the initial discussion resulting from the presentation of practical experiences focused on four key issues: collaboration between creating agencies and archives; methods to ensure early intervention; handling of open, dynamic databases; and preservation scope.

All of the projects presented are subject to a clearly defined legal and regulatory framework, placing demands on the management of public records. This means that agencies are obliged to deliver their records, and hence their databases, to the respective archives, and therefore collaboration between archives and agencies is crucial.

Despite this obligation to deposit records with the state archives, there are limited possibilities for the archives to exercise an influence over the use of specific data formats by the creating agencies. This leads to a variety of approaches. Some of the examples presented indicate that collaboration with agencies is positive and that they are able to impose certain requirements, while others remain subject to the formats, software and standards that the agencies have chosen to use.

[8] The Swiss Federal Archives' Website is at http://www.bundesarchiv.ch/.
[9] Structured Query Language for relational databases

Databases may be delivered in different states to the archives. While most examples were concerned with databases that are closed and no longer in use, the question of current dynamic databases remains to be addressed. In most cases, the archives will deal with periodic snapshots. This issue was addressed in more detail in the second part of the workshop.

While most of the projects deal with the whole workflow of database preservation, some of them focus on a more specific area. The US National Archives' AERIC system serves only the inspection and control processes of dataset preservation. Other projects, while providing general preservation facilities, have a strong specialisation in specific steps of the preservation process, as is the case with the British National Archives' NDAD, which focuses on providing archival description and documentation as well as online access to datasets.

The examples presented highlighted the range of databases that require management and preservation. The NDAD preserves datasets, usually simple tables, predominantly resulting from statistical surveys. Other projects like the Flemish DAVID or the Dutch Testbed deal with more complex databases such as large-scale relational databases. DAVID is the only project that also examines the question of GIS (Geographical Information Systems) preservation, but has not to date been able to present a robust solution for their archiving.

Main Steps

All seven examples presented an overview of the entire preservation process. The main steps involved in this are:

- *Liaison with agencies.* As has already been mentioned, the preservation by state archives requires very close co-operation with the record-producing agencies. The first steps of a preservation process therefore are contacts with agencies, ideally as early as possible in the data creation stages in order to exercise the maximum possible influence over database layouts and formats.
- *Appraisal and selection of data.* This step is closely related to the preceding one and should take place as soon as possible. The prospective appraisal outlined by Thomas Zürcher on the second day of the workshop offers great potential.
- *Conversion* of the data to an archival format and *Accession* into the archives. This takes place in close collaboration with the agencies. Often there are specialised tools to assist data inspection, control and ingestion, like the AERIC facility or the A0 and A1 tools of SIARD that analyse the data structure and assist metadata capture.
- *Description and Documentation.* Description and documentation are closely linked to the accession process. Experiences gathered in the NARA and the NDAD show how crucial good documentation of the context of the database is for future understandability and reusability of the preserved data.
- *Access*. Data will be accessed at a later point in time. This depends heavily on the kind of data, archives, and the respective regulations.

Despite differences in scope and experiences of the projects, all participants agreed on a few common best practices. These include software independence, use of standard formats, and extensive metadata capture. Other common features in practice included the predominance of 'back-end' intervention. For the time being the role of the archives in the choice, introduction and implementation of databases in creating agencies is minimal. Speakers and participants agreed that more involvement in the front end would be highly desirable (and this will in fact be easier in the future), but that there are no certainties as to what extent this will ever be possible.

All seven projects presented at the ERPANET workshop use some kind of migration as part of their basic preservation strategy. It is worth mentioning that no institution favours emulation as an approach to preserve databases.[10] Not only do there not seem to be any existing practical experiences based on emulation, but participants estimate that the perspectives for using emulation for the preservation of database-driven applications are not promising for the future. However, emulation has been mentioned as one possible direction to investigate, and the importance of understanding and documenting the original functionality has been acknowledged.

The solutions presented during the workshop focus on the preservation of the content of a database (data and their structure). The functionality of the system where the data have been created can not be preserved with data conversion and migration. The only way to document how the business processes created data and how the database application supported these processes is to collect system documentation material covering the whole life cycle of the system.

Documentation

Although there was agreement about the importance of documentation, the practical examples presented treated this as a marginal issue, possibly a reflection of its not being the top priority. Participants noted that there is always a danger of insufficient documentation being available, and therefore a tendency to safeguard whatever documentation exists. This documentation is often hybrid, partly on paper, and partly digital. This triggers yet other problems. Digital documentation must itself be processed in order to be preserved in the long term. Hybrid documentations may be united through scanning paper documents and preserving only the digitised version, as with the example of the NDAD.

Potential gaps in the documentation should be located and addressed. One participant recommended the collection of information about the functionality of databases through personal contacts with the creators of the data to access the non-written collective memory as well.

Access and the Role of the Users

Discussions stressed that, if not detailed from the outset, access to preserved databases must at least be theoretically possible or imaginable, if preservation attempts are to be worthwhile. However, the ways the different projects treat access to the preserved databases vary broadly. Two aspects of access came under consideration, both closely linked: the format in which the data are presented; and the skills users must have.

Two distinct strategies are discernible for data formats. While some projects strive to make their preserved databases accessible in as user-friendly a manner as possible, others deliver only the preserved flat files. The best example provided for an access-centred, user-friendly undertaking is the British NDAD. When implementing this project, the main focus was to provide fast and convenient access to digital datasets through a Web interface. This pioneering work allows users to view extensive descriptions of different levels of their datasets and to select the data they want to view. While some basic skills and familiarisation are required, users become quickly acquainted with the system and are able to find the data they need.

[10] The Dutch Testbed project examines emulation approaches, but did not present any approaches related to database preservation.

Another project that intends to facilitate access is the SIARD project of the Swiss Federal Archives. Although still in the planning stage, the way of preserving complex relational database structures through archiving the SQL data definition used to build the database provides for easy reloading of the archived database into a then current relational database system, as long as it has SQL implemented. Thus future users will be able to consult the archived database through a current front end.

At the other end of the scale is NARA, who deliver databases to users wishing to consult them in their preservation format as de-normalised flat files.

Closely related to the questions of managing data access formats are the skills that are expected from users of the data. Participants agreed that this question needs further exploration, taking into account the context of consultation and available assistance. In state archives, preserved records are usually blocked from access for a period of time (unless explicit permission is granted). It was suggested that such issues could be regulated with the aid of access through current database systems.

Extraction, Description, Appraisal, Access and Use

After the first one and a half days' focus on projects and current practice, the rest of the workshop was dedicated to four presentations of specific topics, each covering one step within the preservation process. These papers were a valuable supplement to the first ones. While presenting round-ups on their subjects, they also sharpened participants' eyes for cross-section comparison between the practical projects. Thus they not only achieved a better understanding of them, but contributed also to a general view of the issue, closing the workshop.

Extraction of Data from Dynamic Databases

Many of the databases to be preserved do not exist as closed, complete packages. Frequently databases continue to be in use, with parts of them scheduled to be archived. This is likely to cause problems when databases are concerned, where data are currently being modified.

There are key differences between snapshot databases and temporal databases. While the former represent uniquely the current state, the latter use time stamps to key the data. It was pointed out that snapshots do not indicate when changes occurred, and that therefore certain facts (like several changes between two successive snapshots) can completely disappear.

A first attempt to preserve dynamic databases would be to archive log files together with snapshots. However, since their actual purpose is to assist recovery and auditing they are poorly suited for archiving purposes. In addition to this, archiving log files would pose format challenges. This point was highlighted during discussions, with the possible use of audit tables suggested as an alternative for log files.

Two solutions were proposed. The first consists of archiving all rows of a database that are non-current at a given time, while at the same time deleting them. This, however, means that there is no possibility to view complete time-slices in the archived package. This can be remedied by the second solution, which is to archive snapshots of the whole

current database at specific times, combined with the deletion of non-current records. The frequency of snapshots depends on the frequency of data modification. Before major schema changes or major deletions due to legal or business requirements a snapshot should be extracted. Similar considerations can be made for mixed databases where, for example, master data like information about the persons concerned are held in a snapshot database, whereas different business transactions are linked to them, but forming a temporal database.

The conclusions are mostly evident. Temporal databases are best suited for archiving. Archived snapshots allow for synchronous and diachronous research, but queries may become complex. For other databases neither snapshots nor current archiving fully satisfy all requirements. It is therefore necessary that archivists be involved in database design processes, making sure that fully temporal databases are built, including triggers that write all modifications of the database to an archival database. Further research is needed to deal with schema changes and with partial snapshots. Another issue that should be addressed is the fact that many of today's databases contain references to other databases that are subject to current modifications.

Description of Archival Datasets

During the last decade different archival description standards have been developed, the General International Standard Archival Description (ISAD[G]) being the international framework standard. Recent research at University College London (UCL) into the application of ISAD(G) to archival datasets has led to the development of a commentary and practical guidelines, as well as a new ISAD(G) element set for description of datasets.

The multi-level description rule is key to ISAD(G). It comprises a process from the general to the specific, from the whole archival fonds of an organisation to the item, and allows for macro level descriptions in linked authority files. This rule is suitable for application to datasets. While higher levels of description derive from the administrative function of the database and the lower levels from internal structure (tables, fields, relationships), the middle level (the series level) may not exist and may need to be created artificially. ISAD(G) defines series as "*datasets* arranged in accordance with a filing system or maintained as a unit because they result from the same accumulation, or the same activity… A series is also known as a record series *and when applied to datasets is taken as being related annual accruals of a dataset, or regular snapshots of an accruing system etc.*" According to this definition, UCL distinguished between static datasets, containing data from a finite activity, where 'organic' series can be identified, and active database files, where 'artificial' series must be created. Four types of datasets are presented, specifying the series they yield.

Research has led to the proposal of additional item level description elements for electronic datasets. Besides additions to existing areas, new areas have been proposed or existing ones can be split up. There are, however, further questions and research necessities. The question of 'what is a record in a database?' is of crucial value for the description of databases on the item level. Also, the description of metadata, the separate cataloguing of traditional and electronic archives, and the cataloguing at item level need to be investigated further. Participants pointed out that ISAD(G) is not a very useful tool for the capture of the documentation of a database.

Prospective Appraisal and Pre-archival Metadata Capture of Databases

Prospective appraisal shifts the moment of appraisal from the end of a record's active life to as early a point as possible. As a consequence, the Swiss Federal Archives no longer appraise the records at the moment they are transferred to archival custody, but rather appraise the filing systems of state agencies, examining to what extent a position in the system represents the agency's functions. This prompts, among others, the archiving of entire database systems instead of part.

To facilitate appraisal decisions, a natural approach is recommended: the setting up of a typology of digital information systems (databases and others). The typology adopted by the Swiss Federal Archives uses a top-down oriented approach, taking into account the two fundamental appraisal characteristics, namely a system's evidential and informational value. The evidential value relates to a system's ability to document an agency's course of business and is at its highest if the system is the business function itself, but at its lowest if it is merely a support system. The informational value relates to the information the system can provide about any object or fact of world. This has to be judged from the yet unknown perspective of future researchers, but according to certain criteria. This approach eventually leads to an appraisal matrix with evidential and informational value as the main axes that helps visualise the archival value and facilitates an appraisal decision.

Another prerequisite for prospective appraisal is that a system's metadata be captured at an early stage as well. With the advent of prospective appraisal, description and appraisal are becoming closer to each other. The Swiss Federal Archives' metadata systems include the integration of metadata into the finding system at the creating agencies.

Value-adding, access and use

The Sequence Database Group at the European Bioinformatics Institute[11] is responsible for maintaining and integrating several large-scale nucleotid and protein sequences databases. Nucleotid and protein sequencing today has become a mechanical, low-skill task, generating huge amounts of data (currently around 800 GB per day). The respective databases register around 10 million accesses per day by 100,000 users.

As a result, maintaining these databases must be highly automated to cope with such high transfer volumes. Different tools facilitate the submission of data, the quality check, the distribution of data, and the data exchange with other repositories to make data openly available. As an example, the EMBL Nucleotid Sequence Database was presented. Data are delivered from different institutions, and the database is in constant exchange with other repositories, sequencing projects, and specific databases, and serves as a basis for TrEMBL and Swiss-Prot. Data are stored in flat files, and this format has proven to be stable over the years, despite adjustments and new fields being introduced. XML would not provide additional comfort, but rather demand additional formatting work compared with the current data format. The goals of value adding are a high level of annotation, a minimal redundancy (which is challenging since this kind of data tends to be very redundant), high levels of integration with other databases, completeness and availability. Different levels of service provide different access possibilities. The current file format ensured a certain stability over time, and that only a limited portion of the information was worth retention. It may be cheaper to regenerate the data than preserve them since sequencing has become such a quick technology. However, it may be crucial to record the history of the databases to respond to patent claims and, for example, to be able to document how a database looked and functioned at a given point in time. Databases can be subject to enormous stress while still in active use. Delineating the

[11] http://www.ebi.ac.uk/seqdb/.

spatial scope is particularly difficult due to the tight interconnection between the numerous databases.

Conclusions

Best Practices

During the workshop presentations and discussions a set of best practices emerged. The comparisons between the practical experiences presented, enriched by the additional contributions by participants, indicated that all or most projects shared some common guidelines, rules and approaches. It became possible to grasp a sense that, despite differences, most archives have common strategies and tools to preserve databases.

Conversion and Migration

No project relies on the original bit stream and data format of the databases to be preserved, but all convert to open and stable preservation formats. It is acknowledged that further migration might become necessary as time passes.

Some, but not all, projects store the file they got transferred from the creating agency together with the preservation file in a normalised structure. Security requirements impose the storage of at least one archival copy at a different location.

Adherence to standard formats

The conversion processes conducted by all of the projects discussed aim at reaching a limited range of very basic, open standard formats. These include:

- *Flat files for plain text or tables.* Different kinds of flat files were discussed including fixed-field, fixed-length records, delimited field variable length records, tagged textual documents, and comma separated values. There is a tendency towards using ASCII or EBCDIC encoding, but it has also been pointed out (mainly by the Swiss Federal Archives' SIARD project) that Unicode (UTF-16) should be preferred to allow for multilingualism and special characters.
- *XML for database contents and metadata.* The role and value of XML have frequently been highlighted (for example, the ERPANET Workshop in Urbino, October 2002[12]), especially its high flexibility and understandability and the perspective of a clear exit strategy should the format be superseded in the future. The possibility for XML to be tailored to individual needs is demonstrated by the Norwegian ADDMML, an XML Data Definition Language (DDL) used for metadata capture, and by the Geographic Markup Language (GML) that is being developed to cope with Geographic Information Systems (GIS).
- *Standard Query Language (SQL) for the database structure.* Although every database system vendor adds specific extensions to the SQL definition, this is at the core an international and open standard, defined by the ISO and widely used. The Swiss Federal Archives' SIARD project therefore decided to use standard

[12] http://www.erpanet.org/www/products/urbino/Urbino%20Workshop%20Report.pdf.

SQL (SQL-3) to represent the structure of a database. This allows for the reconstruction of the database for future access.

Software independence

As alluded to above, it is vital to move away from software dependencies. Migration as the basic strategy and the use of standard formats facilitate software independence.

Role of Metadata

The conviction that metadata are absolutely key for digital preservation is firmly established. All projects address this issue as a high priority. Some of the main areas of agreement included:

- Allowing for semi-automated metadata capture through tools. Given the large amounts of data and the number of databases to be preserved, as much automation as possible is a must.
- Allowing for additional manual metadata capture, especially for high-level metadata. High-level metadata are of great importance, and usually there are limited possibilities for automatic capture. Here additional manual efforts are indispensable.
- Including the data producers into the capture of metadata. Particular attention should be paid to the non-written knowledge through communication with database creators and users.
- Storing metadata in a standard format. XML has been considered very promising for fulfilling this task.

Access

The ways in which the different projects provide access or intend to provide it have been widely discussed. According to different backgrounds and environments, access issues have been dealt with in various ways. In fact, it often seems less than clear who is going to access the preserved databases, and what skills s/he will or should have. Agreement was reached, however, on the basic principle that preservation without future access is incomplete and of little use. Even if it is impossible to imagine or view access issues from today's point of view, database preservation must keep all possibilities open so as not to bar future access.

Areas for Further Research

During the workshop it became clear that certain issues provoke intense discussions, while for others there is limited knowledge available. As a result, it is suggested that research efforts be directed in the following areas.

Appraisal of databases

Valuable insight was provided into some of the approaches for appraisal of databases. However, different questions remain open and must be addressed by future research.

Prospective appraisal will play an important role. The tools used to evaluate the evidential and informational value of a database or a database-driven information system need further refinement. In addition, the question to what extent appraisal may be influenced by technical and financial considerations remains open. Further research is needed on how preservation techniques influence the archival value of database born data. As database preservation means extracting and transforming data from database applications which are usually highly structured and thereby limiting the way original users can work with them, investigations are needed on how much evidential value is dependent on a detailed documentation of the system functionalities and the business rules governing creation and use of the data in their original context.

Relationship between databases and records

It has become evident that there is a lack of clarification in defining what needs to be preserved of a database or, in other words, where the records are located in an archival sense. There are several conceivable ways databases and records could be intertwined:

- records are contained, as whole objects, in the database;
- the contents of the database contain records. Each record is spread over tables;
- the contents of the database are the record;
- database data (as whole objects or spread across tables), accessed or presented in a precise manner in the application, form a record;
- the whole database system is the record; or
- a database is not a record at all.

As is easily understandable from this catalogue of possibilities, the identification of the record in a database is crucial for the formulation of preservation requirements. It may imply the preservation of whole database systems, including software facilities, and views in order to preserve the records. Therefore it is an important prerequisite for database preservation to investigate the relationship between databases and records.

Documentation

While it is clearly recognised that extensive, detailed and clear documentation is indispensable for long-term preservation, little thought has been given to recommendations and standards in this area. Most projects use a hybrid of paper and digital documentation. The problems this might cause and the requirements documentation must fulfil have not been fully explored.

Contributions of archival science

On a more general level it is possible to say that the long-term preservation of databases is a collaboration of mainly archival science and information technology. During the workshop presentations it was often remarked that technical solutions have proceeded rather far during recent years, but archival science has not managed to develop a full understanding of the challenges of databases and developed appropriate methods to deal with them. Questions like those above: the appraisal of databases; the relationship between databases and records, and others call for archival discussion and attempts to resolve them.

Participants indicated that until now database preservation has been pervaded with implicit assumptions concerning, among others, suppositions about what should be preserved out of a database, where the records are located, and what access is useful

and desirable. These assumptions often proved to impede comparisons and analyses. It is important that they be investigated, and made transparent to all stakeholders involved as well as the public.

Appendices

Programme

erpaworkshop

Long-term Preservation of Databases
Swiss Federal Archives, Bern, Switzerland
9-11 April 2003

Programme

Wednesday 9th April

09:00	*Registration*	
09:40	Welcome address	Christoph Graf (Director, Swiss Federal Archives)
09:50	Introduction to ERPANET	Seamus Ross (Director, ERPANET)
	SESSION ONE Practical Experiences with Database Preservation	Chair: Niklaus Bütikofer
10:00	Introduction to database preservation	Peter Keller-Marxer (Swiss Federal Archives)
10:45	The French programme CONSTANCE: Twenty years of database archiving	Jean-Pierre Teil (Centre des archives contemporaines)
11:30	Preserving electronic records from database-driven information systems	Filip Boudrez (DAVID)
12:15	*Lunch Break*	
	SESSION TWO Practical Experiences with Database Preservation (2)	Chair: Peter Keller-Marxer
13:30	The Archival Electronic Records Inspection and Control System (AERIC)	Greg LaMotta (United States National Archives)
14:15	Practical Experiences of the Digital Preservation Testbed	Remco Verdegem (Digital Preservation Testbed)
15:00	*Break*	
15:30	The National Digital Archive of Datasets (NDAD)	Kevin Ashley (University of London Computing Centre)
16:15	*End of session*	

19:30 Workshop dinner at restaurant "Bel Etage", Gurten, hosted by Trivadis AG

Thursday 10th April

	SESSION THREE Practical Experiences with Database Preservation (3)	Chair: Seamus Ross
09:00	Tools used for testing and long-term preservation at the National Archives of Norway	Terje Pettersen-Dahl (National Archives, Norway)
10:15	*Break*	
10:40	Software-invariant Archiving of Relational Databases (SIARD)	Stephan Heuscher, Stephan Järmann (Swiss Federal Archives)
11:40	Breakout session	
12:30	*Lunch break*	
13:30	Sum up and recommendations	

	SESSION FOUR Archival methods	Chair: Maria Guercio
14:00	Archiving snapshots or transactions: extracting the right data at the right time from temporal databases	Niklaus Bütikofer (Swiss Federal Archives / ERPANET)
14:30	The Application of ISAD(G) to the Description of Archival Datasets	Elizabeth Shepherd (University College London)
15:15	*Break*	
15:45	Prospective Appraisal and Pre-archival Metadata Capture of Databases	Thomas Zürcher Thrier (Swiss Federal Archives)
16:30	*End of session*	
	Walking tour through the old town	

Friday 11th April

	SESSION FIVE Access, Use, Adding Value	Chair: Hans Hofman
09:00	Value-adding, access, and use: Biological databases as a case study	Rolf Apweiler (European Bioinformatics Institute)
09:50	*Break*	
10:10	Breakout session	
11:10	Sum up, recommendations, conclusions	
12:00	*End of Workshop*	

List of Participants

erpaworkshop

Long-term Preservation of Databases
Swiss Federal Archives, Bern, Switzerland
9-11 April 2003

List of Participants

Kuldar Aas	Estonian National Archives	Estonia
Rolf Apweiler	European Bioinformatics Institute	UK
Andreas Aschenbrenner	ERPANET	Netherlands
Kevin Ashley	National Digital Archive of Datasets	UK
Francisco Barbedo	Arquivo distrital do Porto	Portugal
Ursula Bausenhart	Staatsarchiv Basel-Stadt	Switzerland
Filip Boudrez	DAVID / City Archives of Antwerp	Belgium
Georg Büchler	ERPANET	Switzerland
Niklaus Bütikofer	ERPANET	Switzerland
Montserrat Canela	UNHCR	Switzerland
Daniela Carletti	Discoteca di Stato	Italy
Jean-Marc Comment	Swiss Federal Archives	Switzerland
Jeffrey Darlington	The Public Record Office	UK
Robert de Vroom	ACAM	Netherlands
Louis Faivre d'Arcier	Archives de Paris	France
Sergey Glushakov	Open Society Archives at CEU	Hungary
Heinz Gnehm	Swiss Federal Archives	Switzerland
Mariella Guercio	ERPANET	Italy
Reto Hadorn	SIDOS	Switzerland
Richard Hellinger	UNHCR	Switzerland
Stephan Heuscher	Swiss Federal Archives	Switzerland
Heather Heywood	International Telecommunication Union	Switzerland
Amanda Hill	Manchester Computing	UK
Daniel Hochstrasser	Central Corporate Archives Credit Suisse Group	Switzerland
Barbara Hoen	Landesarchivdirektion Baden-Württemberg	Germany
Hans Hofman	ERPANET	Netherlands
Stefan Järmann	Swiss Federal Archives	Switzerland
Max Kaiser	Austrian National Library	Austria
Lambert Kansy	Staatsarchiv Basel-Stadt	Switzerland

Christian Keitel	State Archives Ludwigsburg	Germany
Peter Keller-Marxer	Swiss Federal Archives	Switzerland
Martin Koerber	Fachhochschule für Technik und Wirtschaft Berlin	Germany
Reto Kromer	Swiss Film Archive	Switzerland
Greg LaMotta	Center for Electronic Records, NARA	USA
Thies Lehmann	Kresta, Schnider & Partners	Switzerland
Keith Lovell	BT Group Archives	UK
Martin Lüdi	Staatsarchiv Aargau	Switzerland
Grant Mitchell	International Federation of Red Cross and Red Crescent Societies	Switzerland
Peter Morgan	Cambrige University Library	UK
Patrick Moser	Staatsarchiv Basel-Landschaft	Switzerland
Pietro Natalo	Discoteca di Stato	Italy
Hege Oulie	Archive, Library and Museum Authority	Norway
Cedric Pauli	Archives communales de Meyrin	Switzerland
Terje Pettersen-Dahl	National Archives	Norway
Marylin Porporato	Archives communales de Meyrin	Switzerland
Bendis Pustina	General Directorate of Archives	Albania
René Quillet	Staatsarchiv Basel-Landschaft	Switzerland
Bill Roberts	Digital Preservation Testbed	Netherlands
Seamus Ross	ERPANET	UK
Martin Rüetschi	scope solutions ag	Switzerland
Raivo Ruusalepp	Estonian Business Archives	Estonia
Lothar Saupe	Generaldirektion der Staatlichen Archive Bayerns	Germany
Claudia Schmucki	Staatsarchiv des Kantons Zürich	Switzerland
Roger Schneider	Swiss Federal Archives	Switzerland
Carolien Schönfeld	Gemeentearchief Amsterdam	Netherlands
Jordi Serra Serra	Arxiu Central, DURSI, Barcelona	Spain
Elizabeth Shepherd	SLAIS/University College London	UK
Jacqueline Slats	Digital Preservation Testbed	Netherlands
Patricia Sleeman	National Digital Archive of Datasets	UK
Gregor Strle	Institute of Ethnomusicology	Slovenia
Jean-Pierre Teil	Centre des Archives contemporaines	France
Remco Verdegem	Digital Preservation Testbed	Netherlands
Matthew Woollard	University of Essex	UK
Jean-Daniel Zeller	Hôpitaux universitaires de Genève	Switzerland
Thomas Zürcher Thrier	Swiss Federal Archives	Switzerland