



Long-term Preservation of Databases

Swiss Federal Archives, Bern, Switzerland

9-11 April 2003

W o r k s h o p D o c u m e n t a t i o n

*erpa*workshop

Long-term Preservation of Databases

Swiss Federal Archives, Bern, Switzerland

9-11 April 2003

Documentation

Contents

Introduction	2
Venue	3
Programme	4
Breakout sessions: themes	6
List of Participants	7
Abstracts of Presentations	9
Speakers' biographies	16

Introduction

Databases as a key technology

Databases have been, and continue to be, a key information technology for several decades, and so their preservation is an urgent issue. The value of information has been sufficiently highlighted, and the importance of databases stems naturally from the importance of information they contain. This is often irreplaceable or could only be recreated at enormous costs, which calls for efficient preservation. Furthermore, beyond their information value databases that are produced to support business processes manifest evidential value that documents these processes.

The different challenges

Not only is the long-term preservation of databases an important and urging problem. According to the context very different tasks and approaches are in the centre of the attention.

In computer science, often the expression “archive” is used in connection with databases. This usually includes backup operations and moving parts of the database into more remote memory. The main aims of such “archiving” operations are data security, freeing storage place, and keeping databases small enough for daily use. It is evident that this use of the term “archive” doesn’t correspond to what is common in archival science. Especially there is rarely a long-term approach to this. Research is being made to expand these archiving functions, approaching them closer to what archival theory requires.

Databases that are established during scientific research share common properties. They are typically very large and represent the results of long and costly research and therefore a considerable scientific and financial value. This is why their preservation is a matter of priority, and discussion focuses on questions of access, re-use, and adding value. Updating and maintaining them, adding further value to them, migrating them to new computing environments should the need arise are issues of central importance, while providing easy and powerful access is a necessary component.

Finally, there is archiving databases in the classical sense of archival theory. It is mainly traditional archives, like state or company archives, that face this challenge. Databases are being produced as parts of business transactions and are valuable both for the information they contain and the evidential value they provide. Just like traditional paper documents they are transferred to an archives where they will be preserved indefinitely. Therefore, the well-known and theoretically firmly underlined archival methods, like appraisal, must meet with more technical issues, such as preservation methods, use of standards, to assure the authenticity and readability of archived databases for the long term. Off-the-shelf solutions to this don’t exist, but many archives and scholars are researching and testing approaches to resolve this urgent problem.

It is evident that the long-term preservation of databases must draw on these most diverse approaches to achieve its goal. Hopefully it will be possible to benefit from different proposals and testbeds to get closer to a solution. This workshop’s principal goal is to gather the information that is available and to foster exchange and communication between the different fields of study. It will ask questions like: How are databases appraised? How do archival science and information technology cooperate for efficient database preservation? What technical solutions are conceivable to preserve databases for the long term? Can much discussed digital preservation strategies like migration or emulation be applied to databases? What best practices (like compliance to accepted standards, meaningful metadata, and

others) should be considered? Are approaches like archiving unused parts of databases relevant from a long-term preservation point of view? Which ones of the solutions developed in theory have been completely or partly implemented and what lessons can be learnt of these efforts?

Areas covered

This *erpaworkshop* will follow two distinct approaches. A strong focus will be the presentation of several practical experiences with database preservation. Seven projects and approaches will be presented, coming from state archives and research projects. They range from well-established solutions, incorporating some years of experience, to current research that is just being put into practice. This first part of the workshop intends to provide participants with an overview on the current state of database preservation. To enable comparison of the various approaches, especially during breakout sessions and conclusions, speakers are encouraged to address in their presentations a range of key questions:

- * In what sectorial and institutional context is the solution used?
- * What underlying preservation policy and/or strategy guide the preservation of databases?
- * How is the whole preservation process from extraction and ingestion to use structured?
- * 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?
- * What are the perspectives for the future?

The workshop's second half will examine the issue from a different point of view. It will present different stages of the preservation of databases, ranging from extraction and description to access and adding value. These presentations will give the opportunity to deepen aspects of database preservation, in particular archival requirements and a scientific data focus on value-adding.

Venue

The workshop will take place at the conference hall of the Swiss Federal Archives in Bern, Switzerland.

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		Chair:
Practical Experiences with Database Preservation		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	<i>Lunch Break</i>	
SESSION TWO		Chair:
Practical Experiences with Database Preservation (2)		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	<i>Break</i>	
15:30	The National Digital Archive of Datasets (NDAD)	Kevin Ashley (University of London Computing Centre)
16:15	<i>End of session</i>	
19:30	<i>Workshop dinner at restaurant "Bel Etage", Gurten, hosted by Trivadis AG</i>	

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	<i>Break</i>	
10:40	Software-invariant Archiving of Relational Databases (SIARD)	Stephan Heuscher, Stephan Järman (Swiss Federal Archives)
11:40	Breakout session	
12:30	<i>Lunch break</i>	
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	<i>Break</i>	
15:45	Prospective Appraisal and Pre-archival Metadata Capture of Databases	Thomas Zürcher Thrier (Swiss Federal Archives)
16:30	<i>End of session</i> <i>Walking tour through the old town</i>	

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	<i>Break</i>	
10:10	Breakout session	
11:10	Sum up, recommendations, conclusions	
12:00	<i>End of Workshop</i>	

Breakout sessions: themes

The workshop presentations will be discussed and supplemented by participants' experience during two breakout sessions. The first of these will address the "Practical Experiences with Database Preservation" sessions, whereas the second will focus on the Archival Methods and Access sessions and provide the opportunity to reach some general conclusions.

To allow for easier discussion participants will be split up in four breakout groups. The discussion subjects will strongly depend on the content of the presentations. As a general idea we propose that the first breakout groups focus on one of the following subjects each:

- What is the role documentation plays in the presented solutions? How well are the preserved databases documented, and what needs to be done additionally?
- How good are the underlying preservation strategies of the presented examples?
- Where are the most convincing approaches for the future? Are there best practices that could be identified?
- How is access to the preserved databases regulated?
- Further questions arising during the workshop.

The second breakout session is invited to discuss the presentations of the workshop's second part, as well as to attempt at a conclusion of the results. Questions that could be examined include:

- A discussion on issues of access and use.
- Who is the future user envisaged to be? What is required of him/her? What background does s/he need to have to access the preserved databases?
- Where are the differences between database preservation and classical archiving? What problems are likely to arise, and how should they be dealt with?
- What recommendations could be derived from the workshop discussions? Is it possible to propose a list of the 5 most important recommendations?
- Further questions arising during the workshop.

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 Bausenhardt	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
Richard Davis	University of London Computer Centre	UK
Robert de Vroom	ACAM	Netherlands
Imma Eramo	Università degli Studi di Bari	Italy
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
Eugene Khodosov	Open Society Institute	Azerbaijan
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	Cambridge 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
Markus Solenthaler	scope solutions ag	Switzerland
Gregor Strle	Institute of Ethnomusicology	Slovenia
Jean-Pierre Teil	Centre des Archives contemporaines	France
Remco Verdegem	Digital Preservation Testbed	Netherlands
Eva Vijupe	Records Management Department, MFA	Latvia
Reto Weiss	Staatsarchiv des Kantons Zürich	Switzerland
Matthew Woollard	University of Essex	UK
Jean-Daniel Zeller	Hôpitaux universitaires de Genève	Switzerland
Thomas Zürcher Thrier	Swiss Federal Archives	Switzerland

Abstracts of Presentations

Jean-Pierre Teil, Centre des Archives Contemporaines

The French Programme CONSTANCE: 20 years of Database Archiving.

As soon as 1978 the French National Archives institution started to think thoroughly to gather and store electronic files. In 1983, a team of four people was working and setting the methodology and the technical bases for archiving digital data files and their relevant meta-data. It was aimed to detect, appraise, collect and store databases from the central French administration. At the end of 1985, a mainframe computer was dedicated to this programme as well as for the electronic management system of the 160 km of shelves for the paper documents kept and communicated by the Centre des Archives Contemporaines in Fontainebleau. An operational team of four archivists and four computer people and system analysts had already treated and stored more than 300 files by the end of 1986.

In the 1994 - 1997 a very major migration occurred: the ministry of culture computer and IT branch who promote standards, hardware and software moved in the then new technology, the machines and operating system UNIX. The Constance team was obliged to migrate 6000 files from the mainframe and EBCDIC technology to the Ascii codification and Unix system. We had to change the hardware, the media and all the data codification. It took 18 months to be secure in the new technology and opened to a broader set of databases and central institutions. Data issued from databases are kept for their historical value, that is forever, or very long term. No software or technical data are kept. The main aim is that future users will get a copy of the files they need, with the relevant documentation and submit these data to the software and computer then operational at this time being.

But, digital data files are part of a broader set of archives that reflect the activities of the issuing institutions. So a second important possibility is that searchers will sometimes not use the data base itself but will conduct a more classical study based on the computerised management system organisation and specificity's rather than on the raw data only.

Filip Boudrez, City Archives of Antwerp/DAVID, Flanders, Belgium

Preserving electronic records from database-driven information systems.

The DAVID-project searches archival solutions for electronic records with archival value created by Flemish institutions and public services. An important criterion for the proposed archival procedures is therefore scalability and applicability within current IT-environments.

Research lead to a separate record-keeping procedure for electronic records created and managed within integrated and dynamic information systems. After all, such information systems and their electronic records require a specific approach. The cornerstones of this approach are identification of records or components with archival value, separation of the different system layers (selection), and a decision-making model. This procedure delivers the answers to the questions of what, how, when and by whom, archiving needs to be done. In this way, it's possible to decide whether all data or only a part of the data, and possibly which functionalities or behaviour, needs to be preserved.

It's important, for the successful application of this archiving procedure, that the information systems are well documented. To decide on the archival appraisal and identification of the components with archival value, the archivist needs to dispose of information on the creation, the management and the archival and technical properties of the records, as well as of

information systems. The conclusion that this information is only seldom guarded systematically in organisations, led to the creation of a new archival instrument: the context inventory. IT-staff, IT-users and the archivist, register metadata on the information system and their records in this context inventory. The information from the context inventory plays an important part in the record-keeping procedure. The metadata are archived along with the electronic records.

This record-keeping procedure aims to preserve electronic records as much as possible in a platform independent way by, among others, using XML-technology and open standards. In most cases this results in migrations to XML at the archiving moment. For this purpose, the data models of databases are converted to DTD's or XML schemes.

Meanwhile, this archiving procedure has been put into practice for the preservation of electronic records with permanent archival value such as the electoral register and the population register. The same approach is used to archive websites and GIS. More information on these archiving procedures and the practical cases, is available on the website of the DAVID-project (<http://www.antwerpen.be/david>).

Greg LaMotta, Archivist, Center for Electronic Records, U.S. National Archives
The Archival Electronic Records Inspection and Control System (AERIC)

The United States National Archives developed the Archival Electronic Records Inspection and Control (AERIC) system to automate the verification process which ensures that federal agencies are transferring the proper electronic files and their supporting documentation, i.e. metadata, to the archives. Verification at the time of transfer is necessary because deficiencies and inaccuracies in the records may be impossible to correct once systems become inactive in the agencies. AERIC matches the data files to their supporting documentation, and then produces reports that show whether the fields in the records contain acceptable and defined values. Most of the documentation transferred to the archives has been on paper causing the entry of data from the documentation, usually the record layouts, to be the most labour-intensive part of the verification process. For this reason, AERIC works best for series of records where few layouts correspond to many files. In cases where many layouts have to be created, the processing archivists often use sampling techniques.

AERIC became operational in 1990, and is constantly being upgraded to process a wider variety of file structures. It can handle most kinds of software independent ASCII and EBCDIC statistical or textual files. AERIC is an Oracle 8i-based application that runs on a Sun Ultra Enterprise 450 server connected to some two dozen client desktop computers. AERIC users are now verifying some 2000 files annually.

Remco Verdegem, Projectmanager, Digital Preservation Testbed, Netherlands
Practical Experiences of the Digital Preservation Testbed

It has been said that the last decennium will be the worst documented era of the 20th century. Information will be recorded exponentially in digital form, without there being reliable methods to preserve this information for the long term.

By the end of 2006 the Dutch Cabinet aims to carry out 65% of its transactions between government and its citizens through digital means. Because of this, there is currently a great deal of work going on to develop strategies, methods, techniques and tools to handle the digital produce of the government in a responsible way.

The most important problem concerning the preservation of authentic digital records is

technological obsolescence. Unless action is taken now, there is no guarantee that current files can be read in future with future technologies.

To research solutions and strategies for this situation, the Ministry of the Interior and Kingdom Relations and the Dutch State Archive Service decided to establish a "Testbed" to gain the essential knowledge and experience.

Approach

The Digital Preservation Testbed is carrying out experiments to establish the best preservation approach or combination of approaches for text documents, spreadsheets, email and databases, applying three different preservation strategies: migration, XML and emulation.

Databases

Digital Preservation Testbed will present a broad-brush overview of some database preservation issues in order to provide a conceptual framework for the main focus of the presentation: bilateral database to XML conversions.

One of the issues is the complex relationship between databases and records. In case of a text document, a spreadsheet or an email message it is clear what the actual digital record is. With a database that relationship is not that unambiguous.

The focus of the Testbed database research is on relational databases. We will investigate the migration of (large) databases, but are currently concentrating on the conversion of databases to XML. We have reviewed several commercial tools available for converting databases to XML, but as we did not have much control over the final form of the XML, we decided to develop our own 'database to XML conversion tool', which will be demonstrated at the workshop.

The use of XML to preserve the content and structure of databases has some obvious advantages: XML is an open standard, widely accepted and applied, is platform and program independent and offers a standardized approach to defining and documenting the structure of the file. Of course there are some drawbacks as well: although XML is human readable, it sometimes is just too much to read. XML and its related standards form a complex material and much pioneering work still needs to be done. For example: what to do with the XML once you have got it? Should you read it back into a future database software package or use standard XML query tools?

Digital Preservation Testbed will address these questions in its research.

Kevin Ashley, University of London Computing Centre
The National Digital Archive of Datasets (NDAD)

NDAD (<http://ndad.ulcc.ac.uk/>) was established in 1997 as a service for the Public Record Office of the UK (now the UK National Archives) whose remit was to accession, preserve, describe and provide public access to government databases selected for permanent preservation. I shall describe some of the mechanisms and procedures we use and contrast them with other approaches. I hope to highlight the strengths and weaknesses of the way we work, and highlight the tensions which all archival systems face between maintaining an authentic record, and providing an easily-accessible resource for reuse.

NDAD's systems are intended to be general, and have to cope with poorly-documented systems as well as well-documented ones, and with data from over 40 years of computing, including many databases which predate the era of the relational database. We must balance preservation needs against the cost of acquisition and preservation, working within a

contract with fixed sums of money and fixed targets for numbers of accessions per year. We took an approach which, whilst well-informed by previous experience in scientific and humanities computing, required development through doing rather than extensive theoretical modelling. We believe NDAD was the first successful service from a national archive which combined the preservation of databases with online access for general users to both their descriptions and their contents. In this respect at least, we feel it was and is a success.

Terje Pettersen-Dahl, National Archives of Norway

Tools used for testing and long-term preservation at the National Archives of Norway

In Norway we differ between registry-based ERM systems and other systems.

For registry-based ERM systems we have a national standard (NOARK) with strict rules for the extraction of a transfer. This gives us an opportunity to develop specific tools for handling this kind of transfers. ArkN3 is designed for testing and presenting these transfers.

Other systems do not have this kind of conformity and requires a different approach. In this case we have also developed a standard, though only for the meta-data. We have defined an XML-DTD called ADDMML (Archival Data Description Markup and Manipulation Language), and next to this DTD we have developed two tools for handling transfers.

The first (Arkadukt) is a form designed to create a description-file following the ADDMML, the other (Arkade) is the tool who performs the actual tests on the transfer, using the ADDMML-file from Arkadukt. Arkade can also – if necessary – do a conversion of the data in the transfer.

A final note: our two standards (NOARK and ADDMML) mostly follow the international standards ISO 15489 and Moreq (NOARK) and ISO 14721 (ADMMML), despite the fact that ours were developed prior to the international ones. We will continually adjust to fully follow the international standards.

Stephan Järmann, Stephan Heuscher, Swiss Federal Archives

Software Invariant Archiving of Relational Databases at the Swiss Federal Archives (SIARD)

SIARD: The Strategy, the workflow and the Software Tools

The goal of the SIARD project was to develop a solution for the preservation of relational databases based on the Swiss Federal Archives' archival strategy.

The workflow is supported by different software tools, which will briefly be demonstrated.

Technical Aspects of SIARD: "SIARD under the hood"

Contrary to the high-level concepts for preserving databases, the focus lies on the technical problems and their solutions implemented in SIARD. The issues discussed range from the technology that was used to implement SIARD to the troubles with standards and their enforcement. The efforts to ease the long-term storage of the archived data and metadata will also be highlighted.

Niklaus Bütikofer, Swiss Federal Archives / ERPANET

Archiving snapshots or transactions: Extracting the right data at the right time from temporal databases

Archiving a “living” database needs decisions about what data should be archived at what point in time since the database is changing constantly. One could wait until the database is closed and then archive the end-state of the data. But, only few databases are closed within a reasonable time span. Most critical databases are used for long periods of time. They are periodically modified in order to cope with changing business needs. Waiting until the database is closed would cause a huge loss of valuable data.

Databases can be separated in two main types regarding their method to represent changes in time: snapshot databases and temporal databases. Where snapshot databases represent only the latest state of data and do not contain information about when the data became valid nor about when they have been acquired by the database, temporal databases record all states of data and / or all modifications of data.

Snapshot databases can only be archived as a sequence of snapshots. The systemlogs, in which the DBMS is writing all modifications, can provide information about the “history” of the data. But they are of limited archival use, because they depend on the specific DBMS. Temporal databases contain the complete history of the data. In theory we could wait with archiving until the database is taken out of use. But, as temporal databases grow very fast in size, and as there are often legal and business regulations requiring deletion in the current system, and as there will be major schema changes from time to time, which can cause information loss, archiving parts of the database will nevertheless be needed.

Best suited for this purpose are combined snapshot–deletion procedures, where the whole database is archived at a given time as a snapshot and all tuples with an end-of-validity or end-of-transaction time before that point in time are deleted. As archived snapshots of temporal databases can be very extensive as well, they may not be taken too often; they should form an appropriate archival “package” in size and time/topic coverage.

Many databases are mixed databases regarding the snapshot and temporal database-types. There, a combined archival procedure using snapshots and extraction of data about completed business transactions will be appropriate.

As only temporal databases can fully be archived in an appropriate manner which is able to satisfy all future user needs, archivists should get involved in the database design process. They should take care that at least the important parts of the database are temporal or that triggers are implemented which write temporal data in a separate archival store every time data are modified in the database.

Further research needs to be done on how schema changes affect the archival process and the archived data. Further clarification is also needed in how to deal with referenced data, which is not in the snapshot or in the archival “package”.

Elizabeth Shepherd, University College London

The Application of ISAD(G) to the Description of Archival Datasets

National and international standards for archival description have been developed and implemented over the past decade for the arrangement and description of historical archives. The international framework standard is the *General International Standard Archival Description* (ISAD(G)) published by the International Council on Archives in 1994, 2nd edition 2000.

Descriptive standards have mainly been applied to archives created and held on traditional media (paper, parchment) but archives are now increasingly created and held in electronic formats, including text documents and databases. The management of electronic records is being addressed by academic projects and by regional groups such as the European Union's DLM-Forum and, of course, ERPANET. Professional groups and national archives are developing standards and guidelines. In the UK, the Public Record Office developed two programmes, one for the management of electronic text-based records of government and secondly, the National Digital Archive of Datasets (NDAD).

The main research for this paper was carried out during a project at the School of Library, Archive and Information Studies at UCL in the summer of 1999, to evaluate ISAD(G) as a framework for the archival description of datasets. The project objectives were to evaluate the appropriateness of the ISAD(G) multi-level rule for the description of datasets, evaluate the scope and relevance of ISAD(G) elements in the description of datasets, identify key omissions in ISAD(G) elements for the description of datasets and compare the use of ISAD(G) for listing datasets with other approaches to listing them. The project aimed to provide archivists with practical cataloguing advice for multi-level description and to give an explanation of some of the key problems that face those cataloguing archival datasets. The Project Manager was Elizabeth Shepherd and the Research Assistant was Charlotte Smith, then completing her MA in Archives and Records Management at UCL. The work has recently been updated to take into account the changes introduced by the second edition of ISAD(G).

This paper is based on the results of the research project. The results comprise a commentary and guidelines on the application of ISAD(G) for the description of datasets. The research focused on the description of datasets at series level and below, since it is at these levels that the special nature of datasets requires particular attention. However, the results also address briefly higher levels of description.

Thomas Zürcher Thrier, Swiss Federal Archives
Prospective Appraisal and Pre-archival Metadata Capture of Databases

The long-term preservation of electronic databases is a costly and laborious task. It is therefore crucial to decide as soon as possible whether a given database should be archived or not. The earlier this decision can be made the greater are the chances that the archives can influence the design and the structure of the database in such a way as to ensure its preservation in the long term. For both paper and electronic records, the Swiss Federal Archives apply the policy of prospective appraisal. This means, that the question whether something is worth to enter the archival custody is answered in the earliest stage as possible. Instead of reviewing the paper records at the moment they are transferred to the archive an appointed appraisal committee reviews the filing system of state agencies before the filing system is used. For electronic records the agencies are required by regulation to cooperate with the archives in the design stage of a new information system.

In order to facilitate the decision about the archival value of a database it is recommended to consider it as an information system rather than a technical artefact. An information system is primarily a socio-technical system that was designed by humans and conceived to achieve a certain goal. The core element of the appraisal policy for electronic records is therefore a typology that focuses on the purpose of the system. The most important one in this context is to give complete proof of the public agencies activity. Further key criteria are the originality of the data and the question whether the database contains entire documents or rather ho-

mogenously structured data or records. On the basis of these criteria a matrix for the archival appraisal of digital information system can be developed.

To appraise a system in the earliest stage of its life cycle requires the capture of the system's metadata as well. The classical archival description must thus be replaced by a pre-archival capturing of metadata, which is compatible with the metadata set used in the archival finding system. Some basic lines of a corresponding metadata model will be outlined.

Rolf Apweiler, European Bioinformatics Institute, Hinxton, Cambridge
Value-adding, access, and use: Biological databases as a case study

In the last few years, the progress in sequencing technology and proteome research has led to a massive rise in available gene and protein sequences from a wide range of species. Structure determination and gene and protein expression data collection is also proceeding at an increasingly rapid rate. In this era of large-scale biology we are not only dealing with experimental molecular biological data but additionally with derived information, acquired with the help of computer programs that can recognise sequence similarities and predict structures and functions. All of this increases the need for elaborate management of the data. The data not only has to be stored in databases but these databases have to be accessible, reliable, and easily manageable. They should contain a minimum of redundancy but a maximum of information. Crossreferencing to other related sources of information is vital as is the comparability of data, which can be achieved by standardisation of terms and data fields. The management of genomics and proteomics data is a difficult undertaking made possible only by the increasingly sophisticated electronic mechanisms available to store, manipulate and communicate information.

The data to be managed covers a wide range of biological information. The core data are the collections of nucleic and amino acid sequences, and protein structures. There are also hundreds of specialised databases available. These range from model organism databases, to protein and gene family databases, resources for experimental proteomics and microarray data and protein-protein interaction databases. Furthermore, this core data can be used to create secondary and integrated databases that combine information from different databases. The journal *Nucleic Acids Research* devotes the first issue of every year, the "Database Issue", to Molecular Biology databases. In 2003, this issue contained articles on 130 biological databases and this is just a fraction of the collections publicly available worldwide.

At the European Bioinformatics Institute (EBI), we develop and maintain a number of biological databases and provide a variety of Bioinformatics tools to facilitate database and similarity searches, and data analysis. In my talk I will provide examples of the core resources maintained at the EBI and highlight the most important issues of database management of such resources with regard to value-adding, access, and use.

Speakers' biographies

Peter Keller-Marxer, born 1964, joined the Swiss Federal Archives (SFA) in 2000. He is the project director of ARELDA (Archiving of Electronic Digital Data and Records), one of the key projects of the E-government strategy of the Swiss government. He is also team leader of the digital archiving staff at the SFA, responsible for the operational area and facilities for day-to-day work in current archiving of digital data and records. Peter Keller-Marxer earned his Ph.D. and M.S. degrees in Theoretical Physics from the University of Bern, where he was working on stochastic computer simulations of theoretical models in quantum spin magnetism.

Jean-Pierre Teil was first employed by the National School of Mines of Paris, in the Computing and Information Technology research centre. In 1983 he integrated the Ministry of Culture to work for the National Archives, in the contemporary archives centre in Fontainebleau (Centre des Archives Contemporaines). He was responsible for setting up the computing centre there in order to manage the 160 kms of shelves, and more specifically to develop and apply the programme CONSTANCE (Conservation et Traitements des Archives Nouvelles Constituées par l'Electronique), which dealt immediately with preserving the first databases. He is now helping and advising the archivists, as well as the electronic document and file producers.

Filip Boudrez studied history at the university of Louvain. In 1997 he obtained the degree of Archivist and Records Manager and the same year he followed training in software engineering and computer programming. In October 2000 he joined the DAVID-project. In the DAVID-project, he develops electronic record-keeping strategies, practical guidelines and best practices for all kind of electronic records (e-mail, office documents, websites, databases, GIS, etc). He also puts the project findings into practice for the City Archives of Antwerp. For this reason, he develops tools for the implementation of record-keeping procedures and new technologies like XML.

Greg LaMotta has been an archivist in the Center for Electronic Records since 1994, when he became responsible for teaching the AERIC system to other staff members and to automated system developers. His knowledge of AERIC was developed within the context of his regular duties of appraising, scheduling, verifying and describing electronic records.

After finishing his studies in Information Technology Remco Verdegem worked for more than 8 years as information analyst at a health insurance company. In October 1998, he joined the Dutch State Archive Service, where he was among others responsible for the functional maintenance of the archival system for paper records. Since October 2000 he is working for the Digital Preservation Testbed (since July 2002 as the project manager), which is sponsored by the Dutch State Archive Service and the Ministry of the Interior and Kingdom Relations.

Kevin Ashley (K.Ashley@ulcc.ac.uk) is head of the Digital Archives Department at the University of London Computer Centre, which operates information and computing services for the UK and European research, education and public sectors. For the past 10 years his group's work has primarily involved the preservation of large-scale digital resources on behalf of other organisations. In many cases this has included providing descriptions of those resources and managing access to them. Most recently these resources have been primarily archival in nature, whether born digital or as digital surrogates, and have involved many types

of information (databases, text, images, video and audio) with different access patterns and cataloguing requirements. The department operates NDAD for the Public Record Office of England and Wales, which deals with government records in the form of structured data. It also operates the National Data Repository at ULCC, which provides digital archiving and distribution services for organisations such as the British Library. In addition, it provides technical infrastructure for integrated archival catalogues such as those of AIM25 (<http://www.aim25.ac.uk/>) and CASBAH (<http://www.casbah.ac.uk/>).

He is a board member of the Digital Preservation Coalition, a member of the Advisory Committee for ERPANET and that of the UK Archives Hub. He speaks frequently on matters related to digital preservation and access and management of digital content, and has also been a contributor to electronic records management training provided by the Archive Skills Consultancy (www.archive-skills.com). His career has previously involved pattern recognition in medical image analysis, network protocol development, standards development, numerical software tools and bar-tending; he has contributed open-source software via organisations such as DECUS for over 20 years.

Terje Pettersen-Dahl holds a Cand. mag. from the University of Oslo, Institute of Informatics (1980). He gathered working experience at the Central bureau of statistics in Norway (National company-register), in different consultancy companies (different systems, mainly programming), at a transport company, and as an independent consultant, before joining the National Archives of Norway, where for the past two years he has been working, among others, on technical standards and digital preservation.

Stephan Heuscher has recently joined the Archival of Electronic Data project at the Swiss Federal Archives as Data Architect. He received his MSc in Information Technology and Electrical Engineering at the Swiss Federal Institute of Technology Zurich in 1999. From 1999 to 2002 he acquired XML and database knowledge by working in several industry projects.

Stephan Järmann studied organisational psychology and computer science at the university of Berne. He works as a research associate for the Swiss Federal Archives since 1996. His tasks include the archiving of electronic records from the Swiss Federal administration, data migration and the development of software tools that support these processes. He is project leader of SIARD (Software Invariant Archiving of Relational Databases).

Niklaus Bütikofer is head of the section information safeguarding in the Swiss Federal Archives. Together with his section he is responsible for inspection and advice in records management within the Swiss government administration and for identification and transfer of all types of records to the archives. Under the project name of ARELDA the section is also developing procedures and tools for archiving electronic records. Niklaus Bütikofer is a member of the committee on current records in electronic environments of the International Council on Archives and he is co-director of ERPANET.

Elizabeth Shepherd qualified as an archivist and worked in archives and records management in local government before becoming a records management specialist at TFPL Ltd. Between 1992 and 2002, she was programme director of the MA in Archives and Records Management at University College London (UCL) and now teaches records management and management skills for archivists on a part-time basis. Her research interests are in the management of electronic records and the development of the archives profession. She serves on the editorial boards of *Archival Science* and the *Records Management Journal*, is a member of the Lord Chancellor's Advisory Council on Public Records, and has recently

published (with Geoffrey Yeo) the book *Managing Records: a handbook of principles and practice* (Facet Publishing, 2003).

Thomas Zürcher Thrier was born 1957 in Basel, Switzerland. He holds an MA in Historical Science and Linguistics (University of Basel) and a Certificate of post-graduate studies in Information Science (Certificat d'études supérieures en information documentaire / University of Geneva). He has been working as an information broker, librarian and archivist for several years. Since 2002 he is a scientific assistant with the ARELDA project of the Swiss Federal Archives where he is responsible for the archival description of electronic systems.

Rolf Apweiler studied Biology in Heidelberg and Bath. He worked three years in drug discovery in the pharmaceutical industry and is involved in the SWISS-PROT project since 1987. He coordinates the SWISS-PROT knowledgebase group at the EBI since 1994, and started the TrEMBL, InterPro, GOA, Proteome analysis and CluSTr projects at the EBI. Since autumn 2001 he is also in charge of the EMBL nucleotide sequence database. Rolf Apweiler is heading a team of currently ~100 biologists and programmers responsible for many proteome and genome related resources at the EBI.

URLs: www.ebi.ac.uk/seqdb/, www.ebi.ac.uk/sprot/, www.ebi.ac.uk/trembl/,
www.ebi.ac.uk/interpro/, www.ebi.ac.uk/proteome/, www.ebi.ac.uk/GOA/,
www.ebi.ac.uk/clustr/, www.ebi.ac.uk/embl/.