briefing paper

digital preservation europe

Database Preservation

Information systems for most organizations are currently supported by databases. Preservation of these databases has to address problems including defining what is to be preserved, the creation and long-term evolution of the preserved objects, organizational support for preservation actions, and technologies that will keep the preserved objects accessible and trustworthy. Some of the issues in database preservation have already been addressed in electronic record preservation, but others result from the specific nature of databases.

Information Systems and Databases

A common architecture for automated information systems distinguishes three layers: the data layer, the logic layer, and the interface layer. The data layer is often implemented as a database. The logic layer includes the business rules and the data processing algorithms. The interface layer deals with presentation and interaction with the users. So, if preserving the look and feel of the screen forms and the reports is a requirement, strategies to preserve information at all three layers are needed. If content and not form is the focus, the data and the logic layers become more relevant. However, in many approaches, preserving data and programs require very different methodologies and techniques. We concentrate here on preserving the databases at the core of the data layer.

Database preservation strategies may also be affected by the nature of data itself. We consider here three main categories of databases: administrative, scientific and document management. Four examples help to clarify the classification. 1 Human resources databases are typical administrative databases. They are mainly concerned with recording facts about the employees. Each fact is recorded in one place, if the database is normalised, but may involve several database records in several tables. 2 Space missions require heavy human and financial investments. The datasets collected by the sensors of the spacecraft and in each experiment performed in the mission are of very high value. Organizing and exploring these datasets and associated context data is the task of scientific databases. 3 A judicial court system stores all the documents related to the processes dealt by the court. These are often digital versions of paper documents or digitally signed electronic documents. The database may contain just the metadata and workflow information or include the documents as well. 4 The information system for a pharmaceutical company combines administrative, scientific and document databases. In addition to the traditional administrative information, pharmaceutical companies, as knowledge intensive organizations, retain a large number of scientific records pertaining to their research and development activities. The strict requirements of the quality system impose complex document workflows.

The success in preserving complex structures like databases depends crucially on the amount of information that is lost in the process and this is inversely related to the amount of metadata included in the preservation package. The metadata must encompass organizational and technical aspects at different levels.

- Contextual information for the database as a whole is required, explaining provenance, function, periods covered, etc.
- There must be a record of technical information about the DBMS, the ingestion process, and the transformations performed.
- A specific component for metadata in database preservation projects is the database schema, including both the table schemas and the set of constraints, essential to capture parts of the meaning of the data.
- Important reports, corresponding to data that is central to main organizational processes may also improve the information preserved.

An issue with increased relevance in database preservation projects is change over time. Databases may stay in operation for several years and the preservation process may start before deactivation. Moreover, information may be deleted from the database because it is no longer useful. A strategy to deal with change can be to periodically ingest a snapshot of the database. A better solution is to embed in the design of the information system an explicit archive step to incrementally add to the preservation.

A View on Standards in Archiving and Recordkeeping

Assuming that we have established what is going to be preserved in the database, a considerable amount of extra information has to be generated. Part of it concerns features of the data that are present in general electronic records, where several existing models and standards are applicable. The "Reference Model for an Open Archival Information System" (OAIS) is an ISO Standard establishing the archival terminology and concepts to be used by all sorts of organizations dealing with archiving issues.

In the context of databases, OAIS requirements have implications in the underlying database, and preferably at the design and implementation phase.
The "General International Standard Archival Description", ISAD(G) covers archival descriptions, prescribing a multi-level organization for the descriptions induced by the provenance of the objects. One of the assumptions in ISAD is that the same set of descriptors is used for items at all levels, resulting in uniform description. Using ISAD, a database may be regarded as a fonds, a table as a series and a table row as a document. If the database preservation form includes a set of tables and their relationships, it will be necessary to describe the context of both the overall object and its parts, using the multi-level uniform description.
MoReq2, the "Model Requirements Specification for the Management of Electronic Records" is aimed at electronic records management systems. MoReq provides principles to guide the institutions implementing electronic records management systems and the system suppliers and developers in what concerns the required functionality and services.
These specifications, and the standards that will possibly emerge, are relevant for database preservation. On one hand, concern with database preservation leads to new requirements in the development of information systems, and the design of databases with preservation in mind; the preservation-oriented requirements in MoReq and similar specifications may inspire these new requirements. On the other hand, a database which has been processed into a preservation form requires management as does any other complex set of records, and therefore these functional specifications also apply.

Some Projects

Several experimental projects have been proposed for exploring the specific aspects of long-term database preservation. A comprehensive survey is available as a result of "The Long-Term Preservation of Databases" ERPANET Workshop in 2003. More recently, the "International Workshop on Database Preservation" held by the National e-Science Centre in Edinburgh has debated the central preservation issues from a databases perspective and reviewed current projects and initiatives.

SIARD

The "Software Invariant Archiving of Relational Databases" (SIARD) is a project developed by the Swiss Federal Archives which has produced a preservation workflow and a set of tools. A separate DPE briefing paper outlines this approach.

The Digital Preservation Testbed

The Digital Preservation Testbed is an initiative of the Dutch National Archives and the Dutch Ministry of the Interior and Kingdom Relations. The approach includes a discussion of the use of several preservation strategies, a structure for the database preservation form and a proposal of concrete actions for database preservation.

Relational Database Preservation through XML Modelling

The ingestion of information from databases is part of the RODA preservation project within the Portuguese National Archives. The main approach includes the transformation of both data and structure into more explicit forms using an XML language and adding EAD description, links to binary files, and a METS package wrapper. The project has produced a database ingest module that follows the OAIS reference model and an interface to support institutions in the process of converting their databases to the preservation form.

Open Issues

Information systems are perhaps the most complex challenge in digital preservation. It is difficult to conceive the consequences of the loss of the huge quantities of information that are currently kept in databases by organizations. For part of this data, we simply expect that the continued interest in its use will provide for their curation. But in many cases the infrastructures that support the databases are fragile, the funding for the institutions or projects that create them are circumstantial and the interest of the public is volatile. There is a host of problems, ranging from seemingly trivial hardware oddities and file formats to operator errors and attacks that threat the electronic records in general.

Further information and resources:

Peter Buneman,Vassilis Christophides, Bertram Ludaescher, Chris Rusbridge, Wang-Chiew Tan and Ken Thibodeau. International Workshop on Database Preservation.National e-Science Centre, Edinburgh, Scotland, 23 March 2007.

CCSDS. Reference Model for an Open Archival Information System (OAIS). Consultative Committee for Space Data Systems, 2002.

Digital Preservation Testbed. From Digital Volatility to Digital Permanence: Preserving Databases. ISBN 9080775819, 2003.

Erpanet Workshop Report. The Long-term Preservation of Databases. ERPANET, Swiss Federal Archives, Berne, Switzerland, 2003.

Stephan Heuscher .Technical Aspects of SIARD. Erpanet, 2003.

IDABC. Model Requirements for the Management of Electronic Records (MOREQ). Cornwell Management Consultants plc (formerly Cornwell Affiliates plc), 2001.

ISAD(G). General International Standard Archival Description, Second edition. 1999.

José Carlos Ramalho, Miguel Ferreira, Luis Faria, and Rui Castro. Relational Database Preservation through XML Modelling, in Extreme Markup Languages, 2007.

David Rosenthal. Engineering Issues in Preserving Large Databases. LOCKSS Program, Stanford University Libraries, 2007.

Cristina Ribeiro - Gabriel David, *DEI-Faculdade de Engenharia, Universidade do Porto and INESC-Porto*, {mcr,gtd}@fe.up.pt