Preservation of databases – Normalized object model

Mårten Stenius, Citrus Oy; and Osmo Palonen; Mikkeli University of Applied Sciences; Mikkeli, Finland

Abstract

Preservation of databases is a complex task that has no evident and straightforward solutions available. At Mikkeli UAS we are researching and evaluating a preservation model called "Normalized Object Model".

We create the Object Model by extracting to an object the structure and metadata of the structure as well as the content and content metadata from databases at preservation. The Normalized Object Model reduces the complexity of the original database by preserving the structure as metadata. The Normalized Object Model relieves maintenance of preserved databases from knowledge on the original structure of them and resists technical quality problems at ingest.

We achieve Normalization by creating one single object structure to contain all database structure elements and one single property structure to contain database content elements. Normalization makes it possible to query several preserved databases simultaneously regardless of structure.

A library of object and property types shall be established to create rules for and assist at planning and designing future, standardized database structures.

Introduction to the database preservation task

Databases are used to store information in all areas of life – science, business, public administration, education, healthcare, publishing etc. Databases are for instance relational databases containing structured information and free text databases containing unstructured information. Relational databases may be used to create hierarchical structures and perform transactions, while free text databases are often used in search engines.

Preservation of databases is a subset of - the preservation of electronic objects. In addition to electronic objects digital preservation also covers archiving of digitized content. The preservation of databases aims at storing and archiving the databases for the use by researchers and others interested. As databases cannot be used or viewed without software and user interface the preservation task will run into problems caused by hardware and software becoming obsolete. These problems must be solved by hardware and software strategies or methodologies in order to maintain the electronic objects.

When deciding on preservation strategies and methods an analysis of the database must be made. As a result of the analysis we will know, what is the goal of the preservation task – to preserve the content, the structure or the functionality of the database.

The Normalized Object Model [1]

The Normalized Object Model is a method, which defines how to extract the structure and information (=metadata) on the structure as well as the content and information on the content from databases that shall be preserved. The central theme in the Normalized Object Model is to reduce the complexity of the original database structure. The reduction is achieved by preserving the structure of the original database as metadata. Future migration efforts will not require knowledge on the original structure of individual databases. The Normalized Object Model is invulnerable to technical quality problems in the ingested or original data. The database owner (holder) must provide the descriptive information on the structure elements.

The object model is originated when we define an object as a (database) structure design describing every structure element of the database to be preserved. The object shall contain descriptive information such as identification, type, name, description, subject words and properties as well as technical information such as size/length, Boolean, date and numeric extensions. Parent and relations may also be described. In addition to objects we define an object property as a (database) structure design describing every content element (field) of the database to be preserved. The object property shall contain descriptive information such as identification, type, name, description, describes object and subject words as well as technical information such as language, character set, value sets (text, date, numeric, Boolean etc.), sorting order. Searchable metadata, privacy, classification level and classification period are special information assigned to properties.

Normalization is created, when the object and property structures are defined to embrace all preserved databases - that is we have one object structure for all database structure elements and one property structure for all database content elements. Normalization makes it possible to query several preserved databases simultaneously independent of individual incompatible database structures. A drawback is that building queries to named properties or objects becomes much more complicated.

Object and property types are used to enhance normalization, by enabling related or differently formatted elements to be assigned the same type instead of having for instance several date properties. Some authority, for instance the National archives, shall be assigned the duty to maintain a common type library. Such a type library is aimed to assist and bring standardization to future database planning and design.

In addition to the object and property metadata, the Normalized Object Model specifies metadata structures for the archive creator, archive (fonds) and series. Archive creator metadata include id, type, name, description, time span, subject words and records management schedule. Archive (fonds) metadata include id, type, name, description, snapshot date, time span, subject words and information system name, description and software names, versions and documentation. Series metadata include id, type, name, description, snapshot date, time span, language character set, subject words, database software name and version, database structure, transfer file type and description.

Why the Normalized Object Model

The Normalized Object Model applies best to such databases where preservation of the content and the structure is important, while preservation of the live functionality is not vital. The

database functionality may be described in appropriate properties. When the live functionality is essential emulation or migration are almost the only alternatives for the preservation of databases.

In the Normalized Object Model we propose the Records Management Schedule to be defined as a formal tool to specify the extraction of structure and content information from the database to be preserved and combining it with schedule information to build the transfer file. The record management schedule must be customized to digital objects and databases especially. When creating the records management schedule for digital objects there are two main objectives: create evidence and appraisal/retention [1, pp 13-20]. To create evidence of a database, requires, that its provenance, the context of creation and its authenticity is documented as well as its processing procedures.

The Normalized Object Model transfer file is a component in the chain creating evidence to preserved databases. The metadata of the transfer file are defined in the records management schedule. Transfer file metadata include id, previous id, media id, type, information system name, creator, metadata specification, coverage, format and file. The transfer file is subject to quality control. Although not the original object the transfer file is preserved to provide evidence of that object.

Several other strategies for the preservation of electronic objects have been proposed during time [1, pp 22-33]:

- Do nothing and technology preservation.

  The main point of this strategy is to receive and store the database as such. This strategy would also imply preservation of the used technology. When the hardware or software required by the database gets obsolete, new and appropriate solutions to the preservation must be created using tools available at that time.

  Strengths:

  The preserved database can execute the original functionality of the database.

  Weakness:

  The hardware stops working before Migration. Knowledge of database structures, functionality, metadata ensuring evidence must be extracted/created, how to make the database available in for instance a National Archive environment including security aspects.

- Emulation.

  Implementing the emulation strategy would create software that runs on a current computer (hardware) making that computer behave identical to the obsolete computer (hardware). The strategy makes it possible to preserve a database in its original form similar to technology preservation. High costs to maintain emulation.

  Strengths:

  Software based, using current hardware. The preserved database can execute the original functionality of the database.

  Weakness:

  High cost to maintain emulation process, Knowledge of database structures, functionality, metadata ensuring evidence must be extracted/created, how to make the database available in for instance a National Archive environment including security aspects.

- Migration.

  There are three applicable Migration methods: Backward compatibility including media replacement, Interoperability, Conversion to standard formats. Database preservation requires a combination of Backward Compatibility and Conversion to standards.

  Strengths:

  Migration to standards reduces work/cost for the following migration cycle. The preserved database can execute the original functionality of the database. Migration can be limited to content migration.

  Weakness:

  Knowledge of database structures, functionality, metadata ensuring evidence must be extracted/created, how to make the database available in for instance a National Archive environment including security aspects.

- Encapsulation.

  Encapsulation is more a framework of preservation elements, where the database and database structure are described and documented thoroughly. The Normalized Object Model is a form of Encapsulation.

  Strengths:

  Knowledge of database structures not required, queries across database enabled, database structure standardization enabled.

  Weakness:

  The database cannot be executed except built queries.

Migration experience

Mikkeli UAS has gained experience of database Migration in the successful Musa migration project reported at last year's conference Archiving 2007 [2][3]. The migrated Mummy Musa radio programme information database conforms to the OAIS Reference model [4]. The migration project was executed in parallel with the research work of the Normalized Object Model [1] and therefore only parts of the model were applicable at the time.

Conclusion

This paper proposes a Normalized Object Model to preserve the content and structure of databases on the long term. The Normalized Object Model solves problems related to maintaining knowledge of individual database structures when the databases are no longer used and become historical objects. It also enables the standardization of the tools and methods to migrate the databases from obsolete software or hardware environments to current environments.

Implementation of the Normalized Object Model specifies requirements on the design and documentation of databases. The design requirements provide tools to standardize database structures and elements on a semantic level as described above. The requirements on database documentation provide methods to understand why and how the structures have been specified such as they are.

The Normalized Object Model proposed in this paper is based on work at the Mikkeli UAS to migrate the Musa database into a normalized SQL database using standard ISO SQL tools only [3] and research results presented in the master thesis of Mårten Stenius [1], where the Normalized Object Model was proposed as a preservation method.

Further research in database preservation including the Normalized Object Model has been initiated through an international research project PAUDEN (long-term Preservation and Access to Uncontrolled Database ENvironments), in which the participating organizations are Luleå University of Technology (Sweden), Mikkeli University of Applied Sciences (Finland), University of Minho (Portugal), National Archives of Estonia (Estonia), National Archives of Finland (Finland), National Archives of Portugal (Portugal).

References

[1] Stenius Mårten: Relaatiotietokantojen pitkäaikaissäilytys Arkistolaitoksessa (Preservation of Relational Databases at the National Archives), Master's Thesis, Helsinki University of Technology, Otaniemi, Espoo, 2006.

[2] ElkaD Loppuraportti, Mikkeli UAS, Mikkeli 2006.

[3] Loponen Mirja, Palonen Osmo: Normalized database preserves radio programme information for internal users and research, IS&T's conference Archiving 2007, the Society for Imaging Science and Technology, Arlington, Virginia, May 21-24.2007.

[4] ConsultativeCommittee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). Washington, DC. 2002..

Author Biography

Mårten Stenius is a Master of Science in Information Technology with long experience of product development and project implementation for the newspaper industry. In the 1990's he managed a team developing an image and text archiving solution for the industry. A few years ago he began research in long time preservation of databases and joined Mikkeli University of Applied Sciences in 2007 to help advance their preservation strategies and services.

Osmo Palonen studies history and IT at the University of Tampere and has made his FKT diploma including graphic arts, IT and management at the Helsinki Institute of Marketing. He worked from the early 1970s to the 1990s in newspaper industry as a journalist, including project and systems management for the editorial IT-systems. Before joining the Mikkeli UAS in 2003 he worked for Honeywell Industrial Automation. Palonen is now in charge of the digital repository, archiving projects and service contracts as well as being the MD of the Disec Oy. He is a member of the board in Union of Business Archives and the Editor of the quarterly archiving specialist magazine Faili.