Practical experiences and challenges preserving administrative databases

Mikko Eräkaski
National Archives of Finland
Rauhankatu 17
PO Box 258, FI-00170 Helsinki
+358 50 363 5769
mikko.erakaski@narc.fi

ABSTRACT
Over the recent years the National Archives of Finland has received databases and registries from various governmental bodies. The information contained in these registries and databases will be transferred to National Archives' long-term preservation service in order to ensure the authenticity, integrity and usability of information over time. This paper introduces how information can be separated from database structures and transferred to archives. The key aspect of preserving database information is comprehensive documentation of extracted data. This is done by applying national SÄHKE2-standard and ADDML-standard developed by Norwegian National Archives.

Categories and Subject Descriptors
H.2 DATABASE MANAGEMENT: H.2.7 Database Administration E.1 DATA STRUCTURES

General Terms
Documentation

Keywords
Long-term preservation, National Archives of Finland, databases, ADDML-standard, legislation

1. INTRODUCTION
The domain of digital preservation is widening from relatively simple documents to more diverse materials such as complex databases, data warehouses, geographical data and research data. Archives need to adopt new kind of methods and techniques in order to preserve and keep databases accessible and trustworthy over time. When dealing with complex databases, it is crucial to cooperate with authorities and other archives.

The Nordic countries have world leading scientists as well as expertise in using administrative personal data in research. This is primarily due to extensive administrative registries and a wide usage of personal identification numbers. Large national databases and registries provide unique source material to study macro-level effects and complex causal questions. Finnish administrative registries are widely used in research today, but in the future, this information may be compromised if not properly preserved for the long term [1].

The National Archives of Finland is facing a major challenge when it comes to preserving databases. Hundreds of registries and databases have been maintained by the public administration during last decades, but only few of them are being maintained steadily for long-term preservation purposes. Our previous experiences have shown that preserving databases is not merely a technical challenge. Public sector organizations also generally underestimate the research value of their databases, which leads to a lack of awareness of preservation issues, appraisal principles, and the appraisal duty of the National Archives.

2. NATIONAL ARCHIVES PRACTICE
The strategy of the National archives is to ensure that its norms and guidelines correspond to the international standards and requirements used in digital preservation. During the past years the National Archives has developed tools and methods based on experiences in other European archives and their database preservation strategies.

The main strategy of the National Archives is to preserve only data, not functionalities or data processing rules and algorithms. Data is extracted from a database system and separated from database structures. Data is stored in XML or CSV format without any software dependent features or binary files. As part of this process all binary files must be extracted from database and converted to suitable format. The National Archives hasn't set strict rules for the form of the data files. Instead the core requirements concern a description of database and obligatory metadata elements. This description is needed in many levels in order to fully understand extracted data and the context of its creation and use. The description of data and transfer to the National Archives is done using standardized SIP structure and metadata. Additional documentation concerning the context, data origin, database management system (DBMS), data models, processing rules and usability guidelines are preserved also in a PDF format. The documentation that needs to be included in SIP has so far been evaluated on a case by case basis.

The SIP structure is defined by the national SÄHKE2-standard, which is an information model, designed for electronic records management systems (ERMS) [2]. The National Archives have developed SÄHKE2 SIP-structure in order to transfer records from diverse electronic records management systems (ERMS) to its long-term preservation service in unified structure. SÄHKE2-structure is also applied in database and registry data, which ensures that all material transferred to National Archives, is transferred in the same structure with similar metadata.

SÄHKE2 metadata is used to describe database and registry data at the collection-level as well as the records-level. SÄHKE-metadata consists mainly of contextual and administrative metadata describing origin, function, information content and possible restrictions. SÄHKE-structure ensures the integrity and permanence of SIP, which is validated automatically in ingestion workflow.

ADDML-standard is used to describe data: tables, fields, variables, codes and their relationships. ADDML is Norwegian Archival Service's standard for technical metadata. ADDML (*Archival Data Description Markup Language)* is used describing a collection of data files organized as flat files. ADDML describes a flat file structure when it is to be exchanged from one system to another. ADDML standard is relatively flexible, which means that it can be adjusted for local requirements and practices for describing different levels of content. This also allows each archive to define its own rules as to how to apply standard [3].

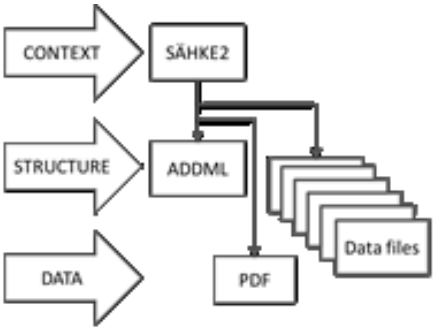

Figure 1. SÄHKE2-standard and ADDML-standard are used to describe context, structure and content of database.

Database data is usually not self-documenting so without sufficient metadata data it can be completely unclear. ADDML describes the meaning of the data by providing a technical and structural data description in a standardized format. It can also describe datasets consisting of more than one table, because it can describe their relationships. The National Archives is cooperating with the Norwegian and Swedish National Archives in order to further develop the ADDML-standard.

3. PRACTICAL CHALLENGES

Databases stay usually in operation for several years and are, in most cases, constantly updated, which causes some challenges for preservation. The first question is whether or not to archive a database if it is still in operation. It is possible to archive snapshots of the entire database at regular intervals or to just archive inactive data, which is no longer modified. The National Archives has exercised the latter method. Most transferred databases have been older databases in which data is no longer altered. The second question concerns documentation. If database has been in operation for longer period it has usually been altered over time. Fields and codes may have changed over time and older data may differ from newer data. Usually database documentation doesn't include information about changes. In the future, organizations should have processes how to keep documentation up to date in the longer term. Also older data is often poorly described. In the case of older registries documentation can be completely lost, available only in paper format, or most likely not up to date. If some parts of the data are obscure then the information can be of no value.

3.1 Complex legislation

In Finland, legislation concerning personal data is complex. The objectives of the Personal Data Act are to safeguard the right to privacy and to promote the development of good processing practice. The Act also regulates destruction and preservation of personal data: *"If a personal data file is no longer necessary for the operations of the controller, it shall be destroyed, unless specific provision have been issued by an Act or by lower-level regulation on the continued storage of the data contained therein or the file is transferred to be archived…"*[4] Public authorities often have problems understanding complex legislation and the Personal Data Act is read to imply that data must be destroyed.

Public authorities are often unaware of appraisal principles, and the appraisal duty of the National Archives, which has a determinative role in the process of the appraisal of public records and data. As a result data in many governmental registers and databases have not yet been appraised by the National Archives, because records creators have not sent their appraisal proposals concerning their registries to the National Archives. Altogether this has led to situations where public authorities have destroyed register data that should be preserved as part of the National Cultural Heritage.

4. REFERENCES

[1] Gissler, Mika & Haukka, Jari: Finnish health and social welfare registers in epidemiological research. Norsk epidemilogi 14/2004.

[2] SÄHKE2-standard: http://www.arkisto.fi/se/saehke2-maeaeraeys (available only in Finnish)

[3] ADDML-standard: http://www.arkivverket.no/arkivverket/Arkivbevaring/Elektronisk-arkivmateriale/Standarder/ADDML

[4] Personal Data Act (523/1999), English translation: http://www.finlex.fi/en/laki/kaannokset/1999/en19990523.pdf