

# Digital Preservation Testbed

## Database Preservation Issues

Remco Verdegem  
Bern, 9 April 2003

# Testbed Background

- Established October 2000 by
  - Ministry of the Interior
  - Ministry of Education, Culture and Sciences
    - Dutch National Archives
- Will finish 1<sup>st</sup> October 2003
- Objective:
  - To secure sustained accessibility to reliable government information in the digital era

# Research Questions

- Advantages of different preservation approaches?
- Factors and effectiveness of each approach?
- Basic Requirements for preservation?
- Which metadata are essential for preservation?
- Options for Attribute preservation?

# Scope

- 4 Archival Record Types:
  - Text documents
  - Spreadsheets
  - Emails
  - Databases
- 3 Preservation Approaches:
  - Migration
  - Emulation
  - XML

# Database



- Three components:
  - Contents
  - Database Management System (DBMS)
  - Application

The database system comprises all three components.

The term database includes at least the contents of a database.

# Different types of databases

- Relational - Oracle, Microsoft Access
- Hierarchical
- Native XML - Tamino
- Object oriented
- Network

# Compared to other recordtypes

- Each database system is unique
- The native environment (application) is not widely available and is generally database specific
- The technical challenge for converting databases into a preservation friendly format is high
- Operational database contents are subject to frequent changes
- The relationship between a record and a database is unclear and is also context dependent.

# Relationship database & record

- 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 is 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
- A database is not a record at all

# What are we trying to preserve?

- The record/s as discrete, easy to access entities.
- The ability to reconstruct the contents of the database.
- The data and input/output screens to form/reproduce records.
- The whole database system.

# Dutch archival regulation

- article 6e

“for databases: the original storage format or ASCII (flat file, with separator tokens), accompanied by documentation, preferably as an XML-DTD, about the structure of the database (at least encompassing the complete logical data model with a description of the entities); queries should be stored in the query language SQL (SQL2)”.

## Our work so far

- Focus on relational databases.
- Conversion of databases to XML, concentrating on content and structure.
- Review of commercial tools for converting databases to XML.
- Design and development of conversion tool.

## XML; pros

- Open standard, widely accepted and applied, well structured.
- Platform and program independent.
- Practical approach to the concept of separation of content, structure and appearance.
- Extensible and controllable; readable by both humans and machines.
- Free - i.e no royalties payable.
- Widely used, so lots of software tools already available.

## XML; cons

- Verbose; indeed human readable, but too much to read!
- XML will be superseded in 5 or 10 years.
- Complex material -> much pioneering work is still to be done.
- What to do with the XML?

### Alternatives:

- Original file format
- ASCII

# XML vs migration

- Migration needs to have both systems running.
- Migration needs to happen every few years.
- Lot of work involved in migration; requires specialist knowledge.
- Conversion to XML can be seen as an intermediate step in a migration between a present day and a future database.
- By converting to XML, you remove the dependency on the present day database system.

## Tool; how it works

- Concentrates on the data in the database.
- Designed to closely follow the structure of the relational database.
- One XML file for each table.
- Separate overview XML file.
- Constraint information in the overview file describes the structure of the database.
- (Optionally) store database views.

# Problems encountered

- JDBC or a ODBC/JDBC bridge
- Details are different for different databases
- Not all of the methods in `java.sql` work in all cases
- Oracle data tables for one application are typically associated with a particular table owner
- Resources
- Images

# Demonstration .....

# Conclusions

- Preservation of databases is still “uncharted territory”.
- Lot of questions, just a few answers.
- XML is able to preserve content and structure of databases.

More information:

<http://www.digitaleduurzaamheid.nl>