



Preservation by Migration to XML

iPres - Beijing - 20071011

Dirk Roorda

Data Archiving and Networked Services

DANS

work on a preservation strategy

- positioning of the XML preservation strategy
- implementing the strategy in software
- pursuing a standard
- with international partners
 - *welcome to MIXED*

Dynamics of preservation

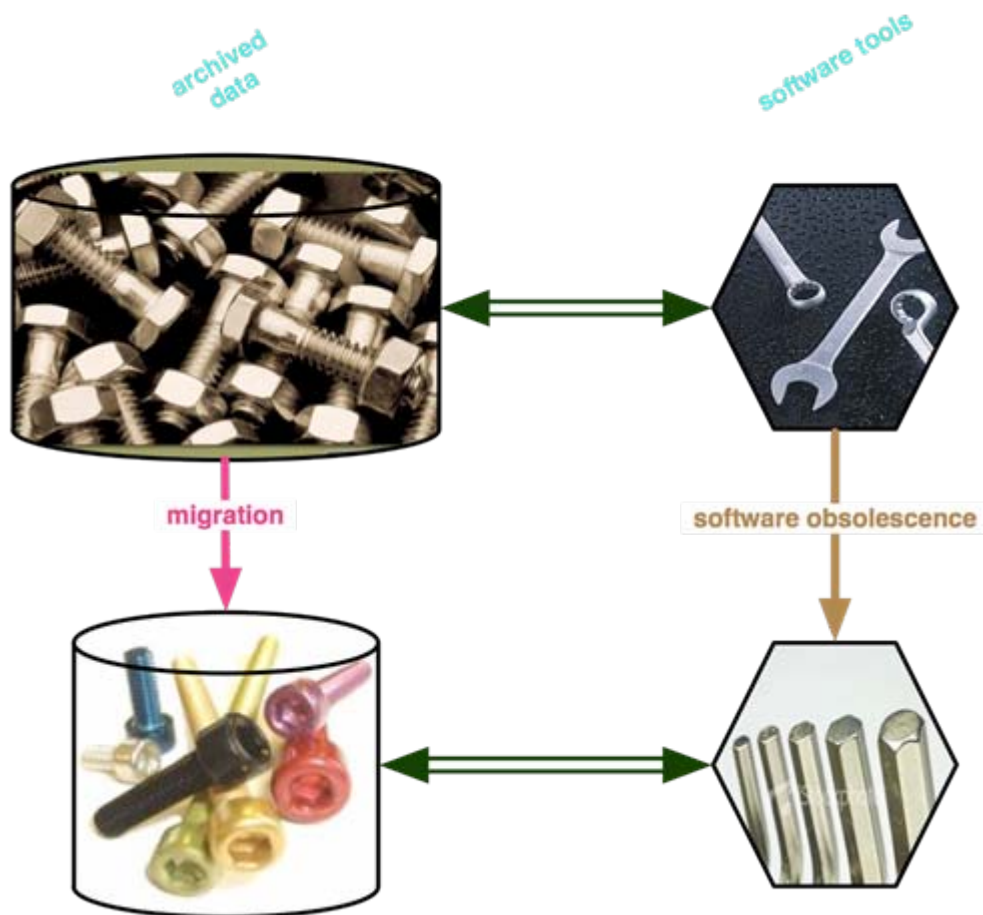
- when the digital context changes
 - emulate: re-implement data and tools
 - migrate: re-represent data, use new tools
- whatever you do, do it smart

To emulate or to migrate?

- archival material = data plus tools
- total recall \Rightarrow emulation
- new use of old data \Rightarrow migration

doing nothing is also an option

Data and tools



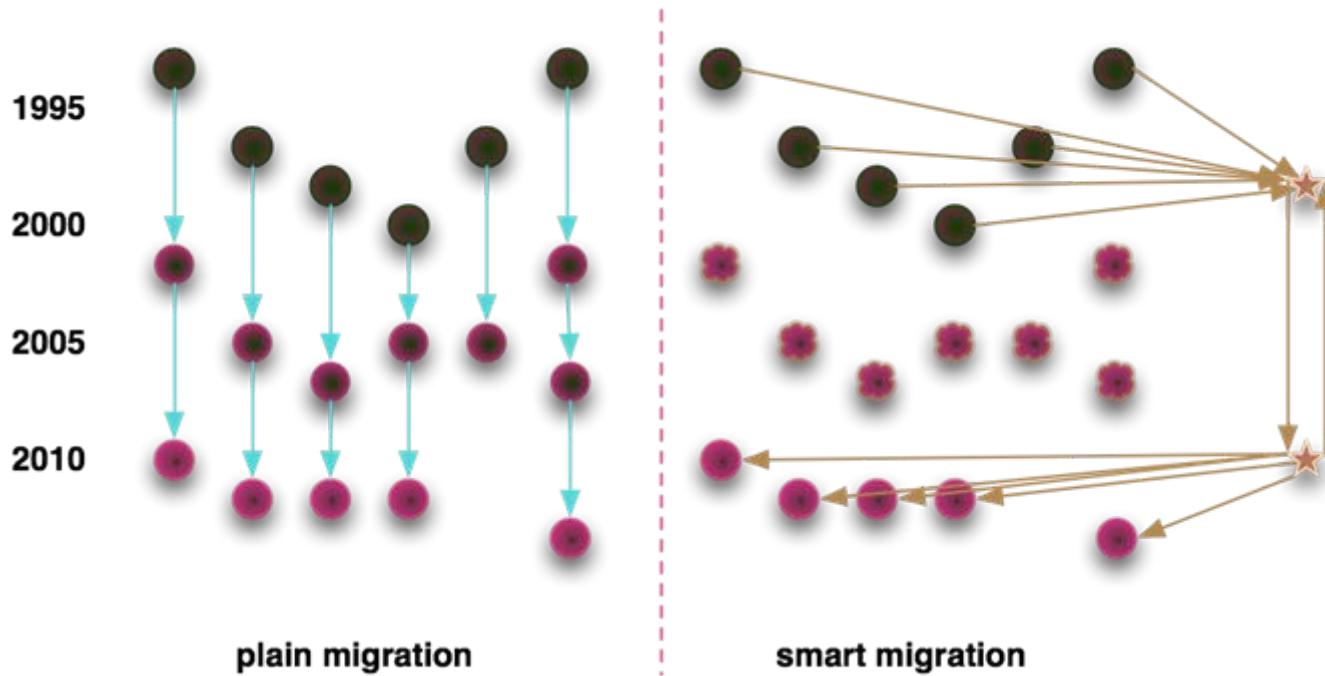
Smart strategies

- multiple related tasks?
- seek *normalization*
 - data representation: NF1-6
 - emulation: universal virtual computer
 - migration?

Diachronic and synchronic

- migration across time (diachronic)
 - original is nearly obsolete
 - original intention might be unclear
- migration within time (synchronic)
 - many different formats
 - vendor specific peculiarities

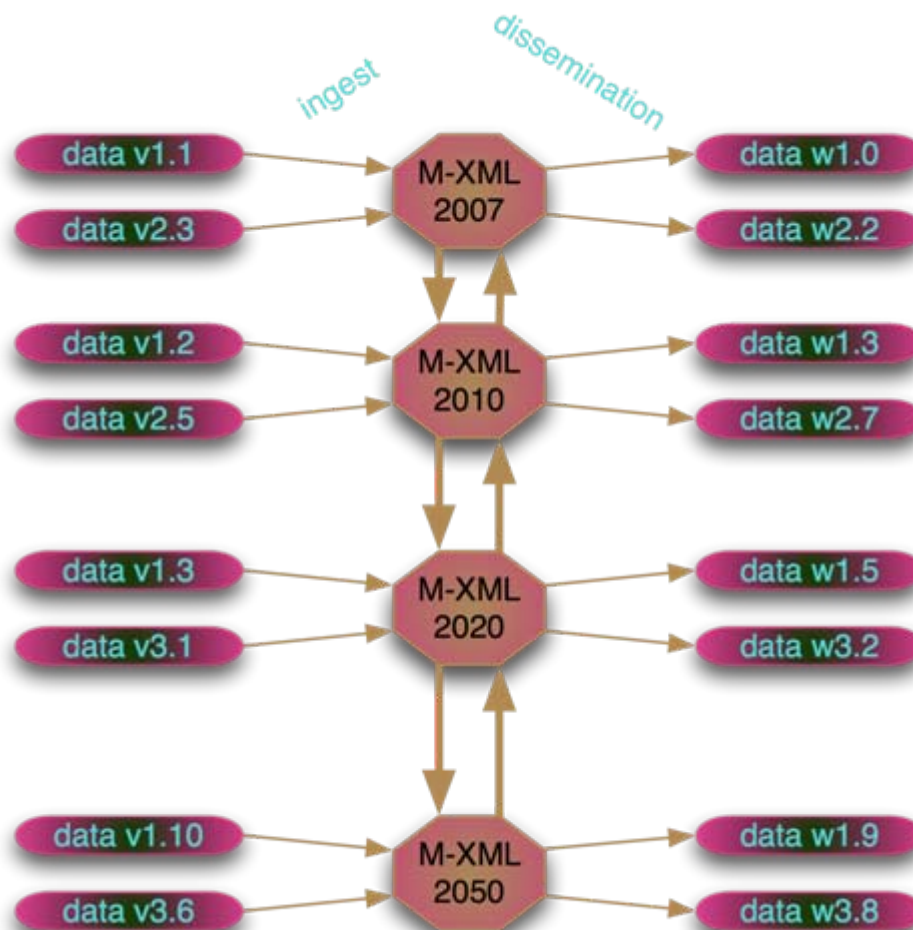
Smart migration



MIXED explained

Migration to
Intermediate
XML for
Electronic
Data

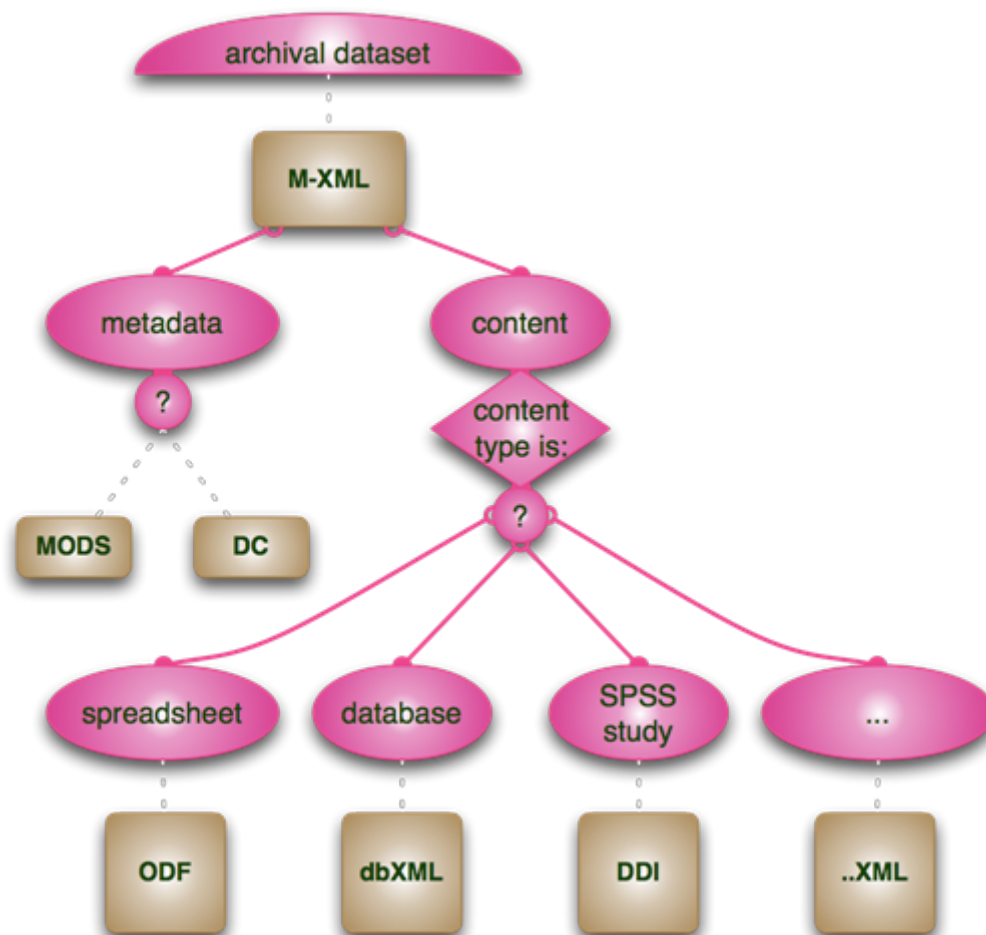
MIXED scenario



Kinds of data

- documents
- spreadsheets
- databases
- statistical data
- images
- word, open(red)office
- excel, openoffice
- access, mysql
- spss, sas
- photoshop, irfanview

Umbrella format



Making it work

- software
 - = framework
 - + modules
- standard
 - = wrapper
 - + metadata
 - + for each kind of data:
 - selected XML standard for that kind

Making software

- make a good product as initial effort
 - generic framework
 - substantial number of conversion plugins
 - for spreadsheets and databases
 - for statistical data
 - connect to repositories, Fedora ready
- integrate efforts of all interested parties by
 - using an open architecture
 - webservices for framework and plugins
 - use third party plugins for SPSS and DDI
 - using an open source paradigm

Helping a standard emerge

- finding a name: Preferred Data Formats for Archives (PDFFA) *suggestions welcome*
- using existing auxiliary standards
- connecting to open source software
- seeking a user base in the archiving world

Using MIXED

- repositories: preservation planning
- interested in file conversion: web services
- individual users: stand alone

Using standards

- XML (Schema, UNICODE)
- OAIS (interface to repositories)
- ODF (spreadsheets)
- SOAP, ESB
- Java, SPRING, OSGI

Co-operation

- DExT (Data Exchange Tools) (UKDA)
- ODaF (Open Data Foundation)
- *you?*

Trends

- end of vendor-specific binary formats in sight
- interchange formats more concerned with preservation

The future ...

- the major data kinds have a preferred preservation format
- the set of preservation formats is standardized
- easy-to-use software converts between preservation formats and custom formats

The end of MIXED?

- if applications can import/export preservation formats directly
- - then no more synchronic conversions
- diachronic conversions keep developing
- - as long as the preservation formats change

Questions ...

- can we standardize preservation formats per data kind?
 - is 1 format per kind sufficient?
- is the time right for an XML standard for data preservation?
- what is the most usable form for the MIXED software?