ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CIRSS

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE

Data Preservation

NSF

INSTITUTE of Museum and Library SERVICES

V4: Standard preservation strategies

Here we review the four classic preservation strategies.

Four common strategies

Replication

Make lots of copies, distribute them widely

Migration

Keep updating your data to new formats, as needed

Emulation

Maintain software that emulates the original processing

Normalization

Convert data sets to a standard format optimized for preservation

Replication pitfalls

[Make lots of copies, distribute them widely]

Ensuring authenticity and identity across copies

Storage costs are non-trivial, depending on

- Scale of data
- Independence of replications
- Number of replications

Each act of replication introduces room for the introduction of

- Errors
- Confusion (same data?)
  - Particularly problematic for changes, slight or large, in encodings)

Most importantly replication does not protect data against technological changes that compromise the viability renderability, (etc) of formats.

Migration pitfalls

[Keep updating your data to new formats, as needed]

Every act of transformation introduces room for:

- Errors
- Context loss
- Information loss

Migration is usually ad hoc,

taking place usually only when needed (e.g, new software tools);
and often in an emergency -- when something critical fails

Migration and migration formats can be uncoordinated and unsystematic, leading sometimes to data sets in multiple poorly understood, lossy, over-specialized, and incompatible formats; and this can lead to confusion about compatibility, and scientific equivalence.

It does not reduce vulnerability to loss (as addressed by replication)

Emulation pitfalls

[Maintain software that emulates the original processing]

Highly expensive, highly complex

How do you identify the properties of any data set (program, etc.) that must be maintained (the *significant properties*).

For which audiences?
Over the long term?

Not all significant properties can be emulated

Does not address the problems addressed by replication or migration

(costs mounting!)

Emulation environments themselves may need to be preserved in turn

Normalization

[Convert data sets to a standard format optimized for preservation]

Here data sets are maintained in a format with standard encodings, syntax, and ontology, and full documentation at all levels.

For simple data sets this may be just documented CSV files.

For more complex data sets XML is commonly used, along with documentation of both syntax and ontology.

OWL and RDF with corresponding serializations are also used.

Data sets in new or specialized formats can be generated as needed

And often a suite of tools for transformations to other syntaxes or ontologies is maintained.

Normalization pitfalls

[Convert data sets to a standard format optimized for preservation]

Some central coordination and support may be required,
or at least a rich culture of open development

Basic protection against loss still required.

Normalization vs Migration

Aspects of normalization (the role of transformations) are similar to migration, but migration (as a preservation strategy) can be distinguished from normalization

Migration creates a *chain* of data sets with the same data in different formats; Normalization is a *hub and spokes* model

We have said that migration formats are ad hoc, uncoordinated, often lossy and over specialized; and can lead to compatibility problems

But perhaps migration is better integrated into the practical reality of creating new formats for existing data than strategies that require transformation in and out of a single core format. (*perhaps*, but . . .).

Transformations

Both migration and normalization strategies involve transforming a data set in one format to a data set in another format,

both presumably with the same information.

Ideally the transformation, as well as the resulting data set, should be documented in a standard computer-processable metadata languages.

Regardless of whether this is a migration scenario or normalization scenario

The next slide indicates one way this could happen.

[This is also an example of *workflow* and *provenance* documentation.]

Example of transformation documentation (very liberally modified from UIUC Medusa record by T. Habing)

<event version="2.1">
<eventIdentifier>
<eventIdentifierType>LOCAL</eventIdentifierType>
<eventIdentifierValue>MEDUSA:b21248fa-75ac-4c45-aae3</eventIdentifierValue></eventIdentifier>
<eventType>MIGRATION</eventType> <eventDateTime>2011-05-03T10:15:32</eventDateTime>
<eventDetail> **The contentdm_record_1.xml file was transformed into the mods_1.xml file using XSLT.** </eventDetail>
<linkingAgentIdentifier>
<linkingAgentIdentifierType>UIUC_NETID</linkingAgentIdentifierType>
<linkingAgentIdentifierValue>UIUC\gnibaht</linkingAgentIdentifierValue></linkingAgentIdentifier>
<linkingAgentIdentifier>
<linkingAgentIdentifierType>FILENAME</linkingAgentIdentifierType>
<agentIdentifierValue> **contentDM_to_MODSv32.xsl** </agentIdentifierValue></linkingAgentIdentifier>
<linkingObjectIdentifier>
<linkingObjectIdentifierType>FILENAME</linkingObjectIdentifierType>
<linkingObjectIdentifierValue> **mods_1.xml** </linkingObjectIdentifierValue></linkingObjectIdentifier>
<linkingObjectIdentifier>
<linkingObjectIdentifierType>FILENAME</linkingObjectIdentifierType>
<linkingObjectIdentifierValue> **contentdm_record_1.xml** </linkingObjectIdentifierValue></linkingObjectIdentifier>
</event>...
<agent version="2.1"> . .

A reformatting recorded in computer processable documentation; here PREMIS XML.

The input and output files are in bold black, the XSL file that specifies the transformation is in red. Documented transformation can support both migration strategy and normalization.

To see how agent is used to represent software associated with a preservation event in this example (search "FIDO") in https://www.loc.gov/standards/premis/v3/sample-records/PREMIS-3-example-1.xml

(The example is liberally modified from UIUC Medusa record)