ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CIRSS

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE

Data Preservation

NSF

INSTITUTE of Museum and Library SERVICES

V1: Introduction

The nature and importance of data preservation (first pass)

Preservation is an important part of curation

A central objective of curation:

efficiently and reliably support the analysis of data,

and enable reuse over time

Despite:

- Physical deterioration or damage to storage devices
- Decay of digital representations (bit rot)
- Changes to file formats or encodings
- Changes to schemas and standards
- Changes to software tools and applications
- Loss or separation of critical contextual information or documentation
- Changes in practices and expectations *and so on*

(adapted from Giaretta, 2011)

A first definition

Digital preservation:

> "… the active management of digital content over time to ensure ongoing access." (US Library of Congress)

Obsolescence

File Format Obsolescence:

- Software upgrades don’t support legacy files
- Format is superseded by another or evolves in complexity
- File format is not compatible with current operating systems
- Company goes out of business or is bought by a competitor

Hardware/Media Obsolescence:

- New, faster computers and storage media
- Decrease in physical size of media (8in to 3.5in floppy disc)
- Reliability and fragility of media

Image from http://www.clir.org/pubs/reports/pub121/sec1.html

Physical Threats

Digital media and hardware are subject to numerous threats that can damage or destroy their readability:

- Environmental (temperature, humidity, light, dust, dirt)
  - Natural disaster (flood, earthquake)
  - Building failure (plumbing, electrical)
- Inevitable hardware failure
- Human error or improper handling
- Sabotage (theft, vandalism)

Context

Examples of metadata that must not be lost:

- File format:
  - Is this ASCII or EBCDIC?
  - Is this TEI version 2.1 or version 5.0?
- Data context
  - Who collected this observations? Where? When? How?
  - What do these attributes and values mean?
  - How were these values calculated?
- Processing
  - What software created this file?
  - What software can read and process (render, perform, etc.) this?

V2: What is data preservation?

The common definitions are misleading

The problem: ***What exactly is preserved?***

Our answer: ***Nothing*** **is preserved**

Our definition: **[you'll see]**

Why the definition is important

Some common definitions of data preservation

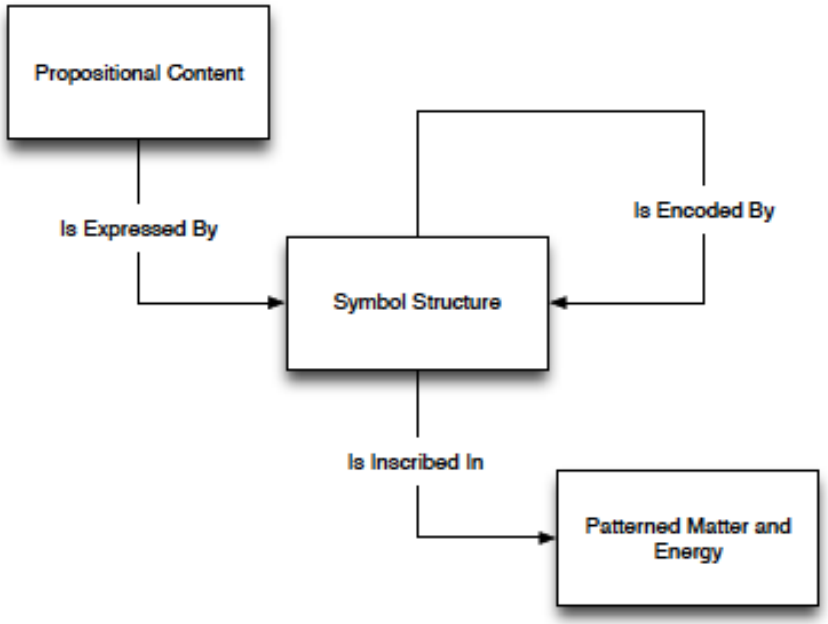
Preservation:

> ". . . The professional discipline of protecting materials by minimizing chemical and physical deterioration and damage to minimize the loss of information and to extend the life of cultural property . . . "
> (Society of American Archivists)

Digital preservation:

> "… the active management of digital content over time to ensure ongoing access." (US Library of Congress)

Recall our data ontology



Now ask yourself:

In a successful preservation scenario, *what exactly is preserved?*

What . . . exactly . . . is . . . preserved?

No, really, ***what?***

The physical thing?

Maybe things in the patterned matter & energy entity type?

That would be actual disks, thumbdrives, hard drives, raid arrays, etc.

Sure, we often do need to take care of these, absolutely

But over time preservation of data (or information) may be successful, even as the individual physical media disappear or become inoperable.

So the continuing existence and integrity of some physical object is not necessary for preservation.

Propositional Content?

Maybe things belonging Proposition Content entity type?

That would be that data or information itself.

e.g., observations, claims, assertions, “facts” etc.

But do those things really need preservation?

Consider this assertion:

100,000 people live in zip code area 61820

Why does its “preservation” require our intervention?*?*

(it can’t crumble, oxidize, dissolve, mold, etc.)

Such things cannot decay, and so cannot be the object of preservation

(at least not in the original and narrow sense of *preservation*).

Symbol Structures?

Maybe things of the Symbol Structure entity type?

But symbol structures also don't need preservation

consider: `up:Z61820 cb:population “100000”`

this is a repeatable sequence of tokens (or characters)
it also cannot oxidize fall apart, dissolve, get moldy, etc

And so it also does not need our help surviving the vicissitudes of nature.

[And in any case:
we often *support preservation* of data by deliberately *changing* symbol structures;
e.g., replacing an obsolete language or format with a different or newer one.]

But if those things then . . . ?

So what is preserved?

We just observed that it can't be any of the things in our data ontology!

The answer is:

Nothing is preserved.

Or perhaps less paradoxically:

No *thing* is preserved.

.

Data preservation is not about preserving data

Preserving physical objects (in the traditional sense of preservation) is part of data preservation, but only part.

> Preservation of physical media must occur for some intervals of time, but data preservation can also take place even when particular media decay — and, indeed, often we deliberately discard media as part of a data preservation strategy.

Preserving propositions (assertions etc) and symbol structures (encodings) is not part of data preservation all: because they cannot decay.

> And, moreover, their continued existence does not ensure preservation.

Preservation is not about preserving any *thing*

Data preservation is not about preserving the existence of objects

It is about *communication with the future*. [1]

The best simple definition is

Ensuring reliable communication with the future [2]

[1] Reagan Moore, Towards a theory of digital preservation, *The International Journal of Digital Curation*, 2008

[2] Simone Sacchi, *What do we mean by preserving digital information'? Towards sound conceptual foundations for digital stewardship* (Doctoral dissertation, University of Illinois at Urbana Champaign), 2015.

The definition expanded. . .

Preservation is not about preservation.

Preservation is *ensuring reliable communication with the future*

More exactly: preservation actions are intended to ensure that future researchers (or other users*)

1) will come into possession of physical media and encodings

2) from which they will correctly recognize the originally intended propositional content

3) and from which they will be justified in believing that this propositional content is in fact the intended propositional content

* this can be adapted to more explicitly accommodate software agents and automatic processing. The key thing is that the process is reliable: all interpretations are (1) correct and (2) justified.

Why is this important?

Because when you emphasize that data preservation is:

Ensuring reliable communication with the future

You see that it is primarily about *understanding* and *credibility*,
not the physical persistence of objects

And that focuses our attention on preservation actions that ensure understanding and credibility, *actions that are too often neglected*.

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CIRSS

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE

Data Preservation

NSF

INSTITUTE of Museum and Library SERVICES

V3: The preservation⇔integration parallels

Preservation goals (again)

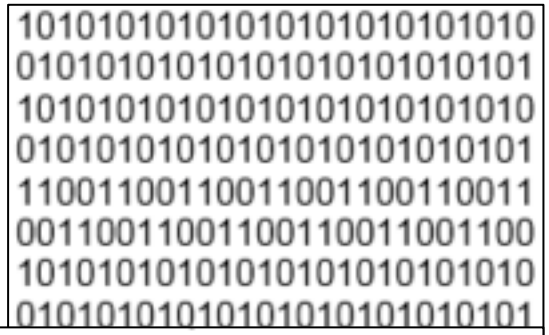
Preservation challenges appear isomorphic to *integration* challenges;
tracing the parallels helps us to better understand
both the challenges and the required interventions.

Preservation goals

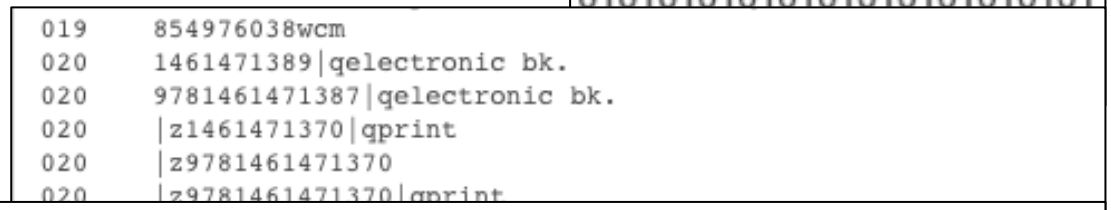
Many organizations and standards
have proposed classifications of preservation goals.

An early statement

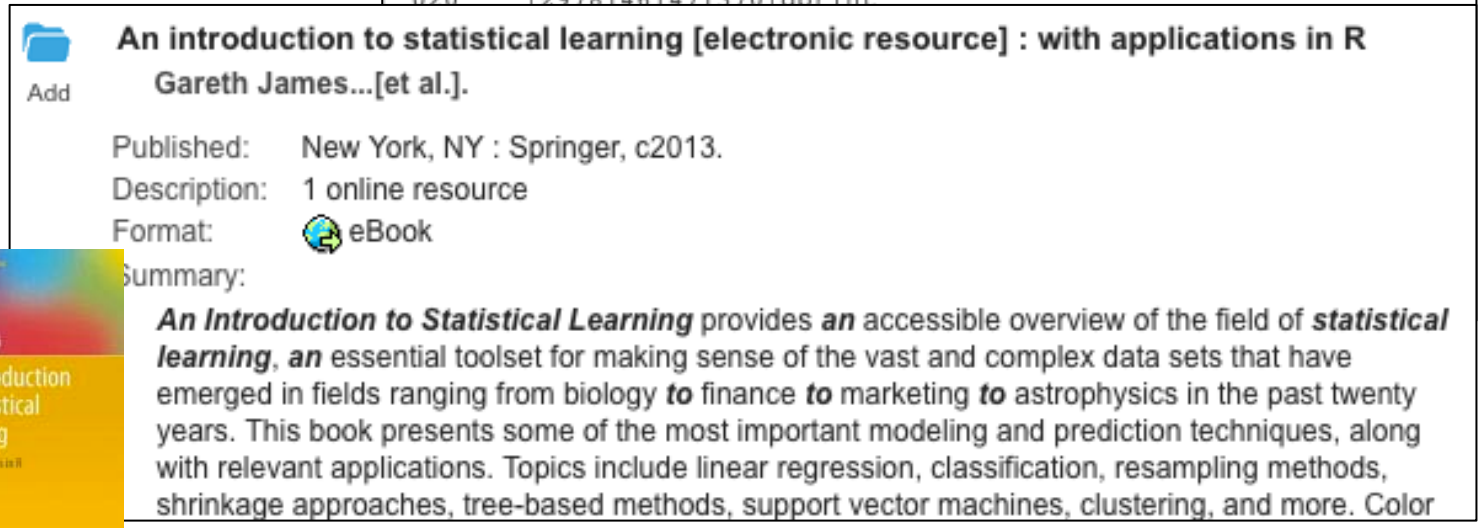
Viable
(can be read from media)



Renderable
(viewable and processable)



Understandable
(interpretable)



See OCLC/RLG Working Group on Preservation Metadata. Image credits: UNC libraries.

And a longer one*

Viable	can be read (correctly) from media
Renderable	can be (correctly) viewed, processed, executed
Understandable	can be (correctly) understood
Authenticatable	can be (correctly) determined to be what it purports to be
Identifiable	can be (correctly) identified and re-identified

And more can be added of course: findable, conformant . . . etc..

Following our own definition of preservation we would emphasize that
each of these five not only must be achievable,
but the user must have *justified confidence* in the result.

*See: PREMIS https://www.loc.gov/standards/premis/

The data integration⇔preservation isomorphism

In many respects the challenges of data *preservation* are the same as the challenges of data *integration*,

only across *time* rather than *space*

.

Let's take a look at data preservation through that lens,
we will be able to apply many things we've already discussed.

(As always, metadata is a key instrument for achieving our objectives)

The role of physical media in communication with the future

Physical objects are essential for *communication with the future* — because we live in a world of space, time, energy, matter, and causality,

And so attending to those physical objects is necessary to *ensure* that communication.

But the individual objects involved need not persist throughout the entire temporal interval:
Communication with the future is achieved via *a chain of overlapping physical objects*

Still, functionality of physical objects must be maintained during relevant intervals in the chain
.

Storage media, and associated hardware (drives, cables, etc.) must be protected against

kinetic damage, chemical decay, electromagnetic insult, etc.,

as well as theft and loss.

For physical media the ensuring of communication is accomplished with
policies and procedures,
the management of associated physical environments;
accurate, complete, and computer readable documentation

[Associated digital objects (APIs, embedded systems, operating systems) must also be preserved.]

Communicating encoding

The decoding of every level of encoding must be reliable.

> From reading 1s and 0s off media, to mapping the bitstream to bytes then integers, then characters or other meaningful symbolic units.
>
> > [With UTF and Unicode some things are easier than they were, but nothing here can be taken for granted.]

As always we ensure the communication of encoding like we ensure all communication:

> through documentation, particularly computer-processable metadata that uses standard vocabularies and a standard serialization syntax.

Usually the most important role of this metadata is to indicate the various data standards being used at different levels of encoding, or to indicate extensions, restrictions, subsetting, or other modifications of standards.

Communicating the syntax

A schema that identifies the structure of the data statements and documents the controlled vocabulary of attributes, identifiers, and values is essential.

This schema should be computer processable so it can be used to validate the structure of the data, to supply data types, and to configure processing tools and applications.

Here again we ensure communication through documentation, particularly computer-processable metadata that uses standard vocabularies and a standard serialization syntax.

And here too the most important role of this metadata is often to indicate the various data standards being used.

Communicating the propositions

This is what it is all about: *Ensuring the reliable communication of what is being asserted*

All the prior attention given to media, encoding, and syntax helps. But there is more to do.

The constituents of propositions are *relationships* and *entities* indicated in the controlled vocabulary for the top level logical syntax. These must be identified, explained, documented.

For this a formal model or ontology is useful.

The schema for this ontology should be computer processable so it can be used for validation, data integration, and management of transformation to alternative syntaxes.

Again we are ensuring communication through documentation, particularly computer-processable metadata with standard vocabularies and standard serialization syntax.

And here too the most important role of this metadata is often to indicate the various data standards being used (here standard ontologies, perhaps modified).

However at this level natural language prose will also be essential:

temperature? *location*? *county*? *race*? species?

. . . such things cannot be explained by formal language alone.

Turtles?

[No, not the RDF serialization language]

As with data integration documentation
there is also the regress problem for data preservation as well.

Metadata for data is itself data,
and must also be preserved . . .
and that will require, among other things, metadata for metadata

Where to stop is a practical matter.

And the foundation will be natural language prose.

This is another reason why existing shared standards are so important.

They already back up mathematical formalities with natural language,
and their existence and use creates communities of shared understanding.

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)

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CIRSS

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE

Data Preservation

NSF

INSTITUTE of Museum and Library SERVICES

V5. Two data preservation standards

OAIS

PREMIS

OAIS

The Open Archival Information System (OAIS) is a model for data archives.*

It is documented *in Reference Model for an Open Archival Information System (OAIS)*, *Recommended Practice*, CCSDS 650.0-M-2 Issue 2, June 2012. https://public.ccsds.org/Pubs/650x0m2.pdf

OAIS was developed by the Consultative Committee for Space Data Systems [CCSDS 650.0-B-2], originally for space agency data

but now the most influential high level data archiving standard.

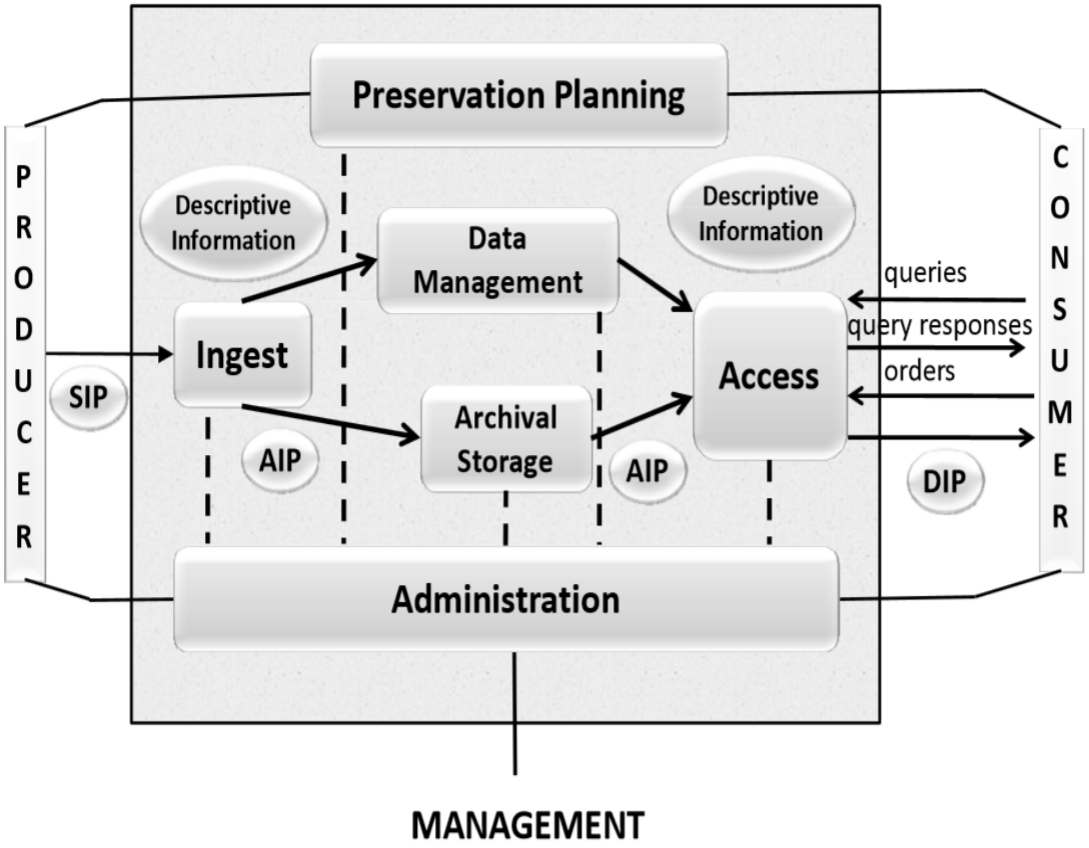
It is very comprehensive, but also very high level,
which means that it is a general framework
but not a specific set of rules, formats, techniques, etc.

Take a look: https://public.ccsds.org/Pubs/650x0m2.pdf

*"OAIS" also refers to archives themselves as well as the standard.

OAIS Functional Model

Section 4.1 – Oasis Functional Entities:

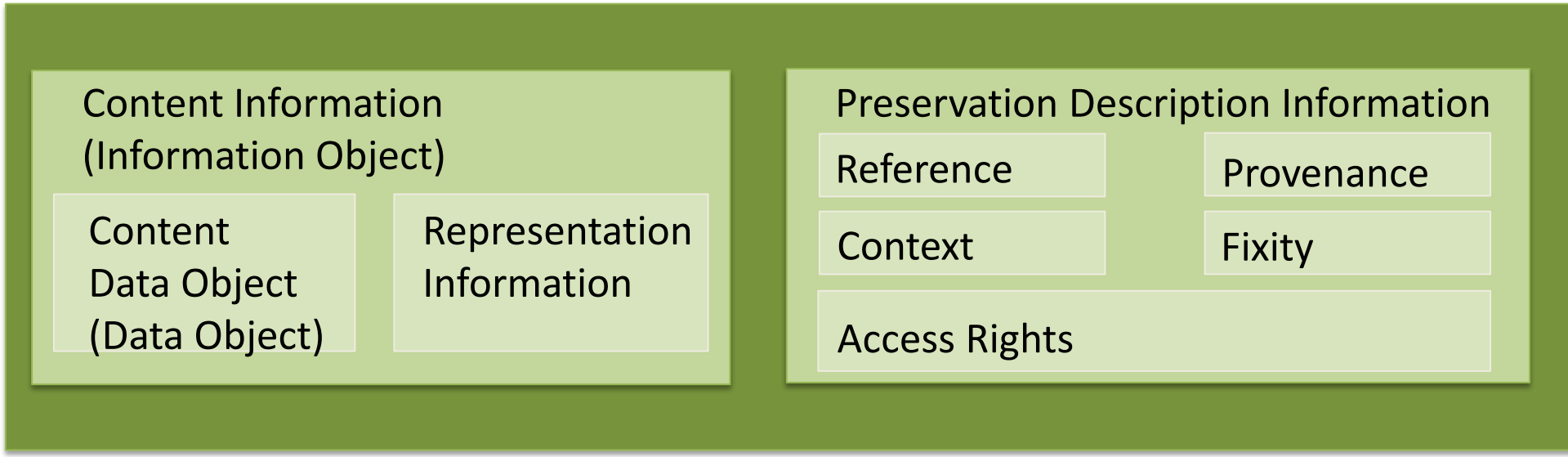


Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2, June 2012

OAIS Information Package

Section 2.2 – OAIS Information Definition

Information Package Definition:



PREMIS

PRESERVATION METADATA MAINTENANCE ACTIVITY

The PREMIS Data Dictionary for Preservation Metadata is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. Developed by an international team of experts, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open-source digital preservation tools and systems. The PREMIS Editorial Committee coordinates revisions and implementation of the standard, which consists of the Data Dictionary, an XML schema, and supporting documentation.

Official Web Site

PREMIS 2.2 (and Schema version 2.3)

- ***PREMIS Data Dictionary for Preservation Metadata, version 2.2*** *(PDF:3.38MB/272pp.)*
  This publication includes the PREMIS Introduction, the Data Dictionary, Special Topics, Methodology and Glossary. The data dictionary and report with supporting documentation are also available as separate documents:
- ***PREMIS Data Dictionary, version 2.2*** *(PDF:1.01MB/227pp.)*
  The PREMIS Data Dictionary for core preservation metadata needed to support the long-term preservation of digital materials.
- ***PREMIS Introduction and Supporting Documentation, version 2.2*** *(PDF:285K/51pp.)*
  Provides information on the background, objectives, data model, implementation and other supporting documentation for the PREMIS Data Dictionary.
- ***PREMIS Data Dictionary Entity Hierarchical Listing, version 2.2***
  A hierarchical list by PREMIS semantic unit.
- **PREMIS Schema 2.2**

News and articles:

- **PREMIS - Implementation Fair, Oct. 6, 2014** NEW!
- **PREMIS - Approved Changes for Version 3.0** NEW!
- **PREMIS - Version 2.3 Now Available** NEW!
- **PREMIS - Schema 2.3** NEW!
- **Preservation Metadata Vocabularies at id.loc.gov** NEW!
- **PREMIS Owl Ontology**
  - **Ontology**
  - **Announcement**
- **PREMIS Implementation Fair at iPRES 2013 in Lisbon, Portugal**

PREMIS Implementors' Group