PREMIS Data Dictionary
for preservation metadata

June 2015 – Revised Novembe

Prepared and maintained by the

PREMIS Editorial Committee

PREMIS Data Dictionary for preservation metadata , June 2015 – Revised Novembe

Available from:

> PREMIS Editorial Committee
> Address line
> Address line
> USA
> 312-606-0722
> Fax: 312-606-0728
> premis@loc.gov
> http://www.loc.gov/standards/premis

ISBN Number

Table of Contents

ACKNOWLEDGMENTS

Special thanks

The following contributed their expertise to previous versions as former members of the PREMIS Editorial Committee:

Steve Bordwell, General Register Office for Scotland

Yair Brama, ExLibris

Olaf Brandt, Koninklijke Bibliotheek, Netherlands

Priscilla Caplan, Florida Center for Library Automation (co-chair of original PREMIS Working Group)

Gerard Clifton, National Library of Australia

Markus Enders, British Library

Noreen Hill, Library and Archives Canada

Karsten Huth, Sächsisches Staatsarchiv, Saxon State Archives

David Lake , U.S . National Archives and Records Administration

Brian Lavoie, OCLC

Yaniv Levi, ExLibris

Bill Leonard, Library and Archives Canada

Rory McLeod, British Library

Robert Sharpe, Preservica

Robert Wolfe, HBO

Zhiwu Xie, Los Alamos National Laboratory

Sally Vermaaten, Statistics New Zealand

Kate Zwaard, U.S. Government Printing Office, Library of Congress

In addition to Editorial Committee members, the following contributed their expertise to the work on the Environment Working Group for PREMIS 3

INTRODUCTION

The PREMIS Data Dictionary is a comprehensive, practical resource for implementing preservation metadata in digital preservation systems. The Data Dictionary defines preservation metadata that

Supports the viability, renderability, understandability, authenticity, and identity of digital objects in a preservation context;

Represents the information most preservation repositories need to know to preserve digital materials over the long term;

Emphasizes "implementable metadata ": rigorously defined, supported by guidelines for creation, management, and use, and oriented toward automated workflows; and,

Embodies technical neutrality: no assumptions made about pres ervation technologies, strategies, metadata storage and management, etc

Attributes

Introduction

Attributes are associated with most of the elements contained in EAD. Attributes reflect named properties of an element and may take on different values, depending on the context in which they occur. In order to set one or more attributes, an encoder should include the name of the attribute(s) within the same angle brackets as the start tag, together with the value(s) to which the attribute(s) is/are to be set. That is, <[tag] [attribute]="[value]"> or <[tag] [attribute1]="[value1]" [attribute2]="[value2]">

For example:
<unitdate unitdatetype="inclusive">1937-1992</unitdate>
or
<unitdate unitdatetype="inclusive" normal="1937/1992">1937-1992 </unitdate>

Most attributes are optional, though some are required. The attribute description indicates whether an attribute is required. This information is also available in the Attributes section of each element description.

The value of attributes may be constrained by the schema using specific attribute type values. For example, @id attribute is of type ID, which constrains its value to a string beginning with an alphabetic character. An @id value must be unique within the EAD instance within which it occurs, that is, no other tag in the entire document can have the same @id value. EAD attributes have the following data types:

anyURI:
A Uniform Resource Identifier. This may be a Uniform Resource Locator (URL) or a Uniform Resource Name (URN). Both relative and absolute URIs are allowed.

ENTITY:
The name of a nonparsed entity that has been declared in the declaration subset of the document. For example, @entityref must contain the name of an entity that has been declared in the declaration subset. Processing software can use the reference to the nonparsed entity to display the entity in the body of the text or in a new window.

ID:
Unique identifier. For example, most elements have an @id, so that a unique code can be established for and used to refer to that specific element. The content of the @id is of the type called "ID". Parsers verify that the value of attributes of type "ID" are unique. The values of @id must begin with an alpha, not numeric, character, either upper or lowercase, and may contain a . (period), : (colon), - (hyphen), or _ (underscore), but not a blank space. See also attributes of type "IDREF."

Capitalization of data types follows the documentation found in the W3C Recommendation XML Schema Part 2: Datatypes Second Edition (http://www.w3.org/TR/xmlschema-2/).

IDREF:
ID reference value; must match an existing ID of another element in the document. For example, the <ptr> element has a @target attribute that can only be an "IDREF," which means it has to reference a valid ID in another element.

IDREFS:
List of ID reference values.

NMTOKEN:
A name token, which can consist of any alpha or numeric character, as well as a . (period), : (colon), - (hyphen), or _ (underscore), but not a blank space. A number of attributes in EAD where a character string from a code list is to be used are of the type "NMTOKEN".

string:
The most general data type, a string can contain any sequence of characters allowed in XML. Certain characters may have to be represented with an entity reference, for example < for <, and & for &.

token:
A type of string that may not contain carriage return, line feed or tab characters, leading or trailing spaces, and any internal sequence of two or more spaces.

The attribute value definitions in the DTD versions of EAD3 differ slightly from those of the Relax NG and W3C Schema versions. The DTD has a limited set of attribute types so the anyURI, token, and string data types were converted to "CDATA" (i.e. Character Data).

When the EAD schema limits attribute values to a few choices, those values are declared in the schema in what is known as a "closed list." For example, the values of @audience are limited to either "external" or "internal." Other attributes are associated with semi-closed lists. Such lists include those values believed to be the most useful in many contexts, but other values are allowed. For example, <dsc> defines several values for @dsctype, including "otherdsctype" which may be used with @otherdsctype to specify values that are not in the semi-closed list for @dsctype. The definitions for some values in the closed and semi-closed lists appear below.

The following is a complete list of all the attributes that occur in EAD, and some discussion of how they may be used. Further, context-specific information about the use of certain attributes may be found in the "Attribute usage" section of the element descriptions.

@abbr Abbrevation

Summary:	An abbreviation for a word or phrase that is expressed in an expanded form in the text of the current element; used for searching and indexing purposes. Available only in <expan>.
Data Type:	token

Elements

<abbr>

<abbr> Abbreviation (Table of Contents)

Summary:	An element for encoding the shortened form of a word or phrase.
May contain:	[text]
May occur within:	abstract, addressline, archref, author, bibref, citation, container, conventiondeclaration, date, datesingle, didnote, dimensions, edition, emph, entry, event, fromdate, head, head01, head02, head03, item, label, localtypedeclaration, materialspec, num, p, part, physdesc, physfacet, physloc, publisher, quote, ref, sponsor, subtitle, titleproper, todate, unitdate, unitid, unittitle

Attributes:		
	altrender	Optional
	audience	Optional (values limited to: external, internal)
	expan	Optional
	id	Optional
	lang	Optional
	script	Optional

Description and Usage: Used to record the abbreviated form of a word or phrase, for example, an acronym.

Use <abbr> within <conventiondeclaration> to identify the code for a thesaurus, controlled vocabulary, or another standard used in creating the EAD description. To improve interoperability, it is recommended that the value be selected from an authorized list of codes such as the MARC Description Convention Source Codes (http://www.loc.gov/standards/sourcelist/descriptive-conventions.html).

In other elements, use <abbr> with @expan to encode abbreviations as they occur within the description, if you wish to use an abbreviation while also providing its fuller form.

Availability: Within <conventiondeclaration>: Optional, not repeatable

Within other elements: Optional, repeatable

Examples:
```
<conventiondeclaration>
    <abbr> ISAD(G) </abbr>
    <citation>ISAD(G): General International Standard
    Archival Description, second edition, Ottawa
    2000</citation>
</conventiondeclaration>
<didnote>File also contains materials from the
```

```
            <abbr expan="American Civil Liberties Union">
            ACLU </abbr>
        </didnote>
        <c02>
            <did>
                <unittitle>
                    <abbr expan="United Nations Educational,
                    Scientific and Cultural Organization"> UNESCO
                    </abbr>
                </unittitle>
            [. . .] </did>
        </c02>
```

<objectIdentifier>

Semantic unit objectIdentifier (Table of Contents)

Semantic components:	objectIdentifierType, objectIdentifierValue
Definition:	A designation used to identify the Object uniquely within the preservation repository system in which it is stored.
Rationale:	Description you can use <to_get_anglebrackets>
	Each Object held in the preservation repository must have a unique identifier to allow other entities to refer to it and to relate it to descriptive, technical, and other metadata unambiguously.
Data Type:	Container.
Description and Usage:	**Intellectual Entity / Representation**
	Applicability Applicable Obligation Mandatory Repeatability Repeatable Example IE / Representation example
	PREMIS File
	Applicability Applicable Obligation Mandatory Repeatability Repeatable Example File example
	PREMIS Bitstream
	Applicability Applicable Obligation Mandatory Repeatability Repeatable Example Bitstream example
Creation/Maintenance notes:	An identifier may be created by the repository system at the time of ingest, or it may be created or assigned outside of the repository and submitted with an object as metadata. Similarly, identifiers can be generated automatically or manually.
Usage notes:	The objectIdentifier is mandatory for all Objects stored.
	The objectIdentifier is repeatable in order to allow both repository assigned and externally-assigned identifiers to be recorded. See "Creation/Maintenance" note above.
	Primary identifiers must be unique within the repository. They may be preexisting, and in use in other digital object management systems. Ideally, secondary identifiers should also be unique but sometimes this is not possible (e.g., if the values are inherited from a legacy system which did not enforce this or only identified items at a higher level). Identifiers for each item must be sufficient to identify the item uniquely at the appropriate level of aggregation. For example, an Intellectual Entity that represents all books

in the same edition could use an ISBN but this would be insufficient to identify a particular copy of that book.

A preservation repository needs to know both the type of object identifier and the value. If the value itself contains the identifier type (e.g., "oai:lib.uchicago.edu:1"), the identifier type does not need to be recorded explicitly. Similarly, if the repository uses only one type of identifier, the type can be assumed and does not need to be recorded.

<objectIdentifierType>

Semantic unit objectIdentifierType (Table of Contents)

Definition: A designation of the domain within which the object
                                 identifier is unique.

Rationale: Description you can use <anglebrackets>

                                 Identifier values cannot be assumed to be unique across
                                 domains; the combination of <objectIdentifierType> and
                                 <objectIdentifierValue> should ensure uniqueness.

Data Type: Value should be taken from a controlled
                                 vocabulary. A controlled vocabulary is available at
                                 http://id.loc.gov/vocabulary/identifiers.html.

Description and Usage: **Intellectual Entity / Representation**

                                 Applicability Applicable Obligation Mandatory Repeatability
                                 Not-repeatable

Examples

1	ISBN (Intellectual Entity)
2	DOI (Intellectual Entity)
3	DLC
4	DRS
5	hdl:4263537

PREMIS File

Applicability Applicable Obligation Mandatory Repeatability
Not-repeatable

Examples

1	DLC
2	DRS
3	hdl:4263537

PREMIS Bitstream

Applicability Applicable Obligation Mandatory Repeatability
Repeatable

Examples

1	DLC
2	DRS
3	hdl:4263537

Usage notes: The type of the identifier may be implicit within the repository as long as it can be explicitly communicated when the digital object is disseminated outside of it.

<objectIdentifierValue>

Semantic unit objectIdentifierValue (Table of Contents)

Definition: The value of the <objectIdentifier>.

Data Type: None

Description and Usage: **Intellectual Entity / Representation**

Applicability Applicable Obligation Mandatory Repeatability Not-repeatable

Examples

1	0 00 221804-6
2	0000000312

PREMIS File

Applicability Applicable Obligation Mandatory Repeatability Not-repeatable

Examples

1	IU2440
2	WAC1943.56
3	AMNH
4	CD269/ CD269/70/10
5	1001/dig/ pres/2004-024
6	http:// nrs.harvard.edu/ urn-3:FHCL.Loeb:sa1

PREMIS Bitstream

Applicability Applicable Obligation Mandatory Repeatability Repeatable

Examples

1	IU2440-1
2	IU2440-2

Semantic unit objectCategory (Table of Contents)

Definition:	The category of object to which the metadata applies.
Rationale:	Description you can use <anglebrackets>
	Preservation repositories are likely to treat different categories of objects (Intellectual Entities, Representations, Files, and Bitstreams) differently in terms of metadata and data management functions, it is therefore important to differentiate between the categories.
Data Type:	Value should be taken from a controlled vocabulary. A controlled vocabulary is available at http://id.loc.gov/vocabulary/preservation/objectCategory.html .

Description and Usage:

Intellectual Entity / Representation

Applicability Applicable Obligation Mandatory Repeatability Not-repeatable

Examples

1   intellectual entity
                                                               representation

PREMIS File

Applicability Applicable Obligation Mandatory Repeatability Not-repeatable

Examples

1   file

PREMIS Bitstream

Applicability Applicable Obligation Mandatory Repeatability Not-repeatable

Examples

1   bitstream

Usage notes: A filestream should be considered a file.

Semantic unit preservationLevel (Table of Contents)

Semantic components:	preservationLevelType, preservationLevelValue, preservationLevelRole, preservationLevelRationale and preservationLevelDateAssigned
Definition:	Information indicating the decision or policy on the set of preservation functions to be applied to an object and the context in which the decision or policy was made.
Rationale:	Description you can use <anglebrackets>

Some preservation repositories will offer multiple preservation options depending on factors such as the value or uniqueness of the material, the "preservability" of the format, the amount the customer is willing to pay, etc. In such circumstances the <preservationLevelValue> that applies may need to be directly associated with an Object. The choice of a particular preservation option for an object may also require further explanation. This can depend on the preservation functions expected to be applied to the object (which can be described by assigning a <preservationLevelType>) and/or the context in which a set of preservation options is applicable (which can be described by assigning a <preservationLevelRole>). The distinction between <preservationLevelType> and <preservationLevelRole> can be illustrated by examples. One possible preservation level type is "Bit preservation level". This might have values of 'Low', 'Medium' or 'High', where, for example, in 2015 technology examples for:

- 'Low' means ordinary on-site backup
- 'Medium' means two copies on different media types with a minimum of 150 km distance between the stored copies, with separate checksums that are integrity checked annually
- 'High' means solutions a minimum of 5 independent copies on a variety of storage media distributed over different organizations in several continents with quarterly integrity checks.

The <preservationLevelRole> can then be used to distinguish, for example, if this level is an aim or has actually been achieved. It is also possible to assign the rationale for choosing the value (which can be described by adding a <preservationLevelRationale>).

Data Type:	Container

Description and Usage:	**Intellectual Entity / Representation**

Applicability Applicable Obligation Optional Repeatability Repeatable

PREMIS File

Applicability Applicable Obligation Optional Repeatability Repeatable

PREMIS Bitstream

Applicability Non-applicable Obligation n/a Repeatability n/a

Creation/Maintenance notes:	The preservation level may be assigned by the repository or requested by the depositor and submitted as metadata. The repository may also choose to record additional metadata indicating the context for the assignment of the preservation level.
Usage notes:	If the repository offers only a single preservation level or the preservation level can be calculated externally (e.g., based on the information in a technical registry or by the type of collection), this value does not need to be explicitly recorded with Objects within the repository. Application of a particular set of <preservationLevel> semantic units may only cover a single Representation of an ob ject: Representations in other technical forms or serving other functions may have a different <preservationLevel> applied. The container may be repeated if a preservation level value needs to be recorded in additional contexts (see <preservationLevelRole>, page 24) or part of a context (see preservationLevelType, page 22).

Elements

<abbr> Abbreviation (Table of Contents)

Summary:	An element for encoding the shortened form of a word or phrase.
May contain:	[text]
May occur within:	abstract, addressline, archref, author, bibref, citation, container, conventiondeclaration, date, datesingle, didnote, dimensions, edition, emph, entry, event, fromdate, head, head01, head02, head03, item, label, localtypedeclaration, materialspec, num, p, part, physdesc, physfacet, physloc, publisher, quote, ref, sponsor, subtitle, titleproper, todate, unitdate, unitid, unittitle

Attributes:		
	altrender	Optional
	audience	Optional (values limited to: external, internal)
	expan	Optional
	id	Optional
	lang	Optional
	script	Optional

Description and Usage: Used to record the abbreviated form of a word or phrase, for example, an acronym.

Use <abbr> within <conventiondeclaration> to identify the code for a thesaurus, controlled vocabulary, or another standard used in creating the EAD description. To improve interoperability, it is recommended that the value be selected from an authorized list of codes such as the MARC Description Convention Source Codes (http://www.loc.gov/standards/sourcelist/descriptive-conventions.html).

In other elements, use <abbr> with @expan to encode abbreviations as they occur within the description, if you wish to use an abbreviation while also providing its fuller form.

Availability: Within <conventiondeclaration>: Optional, not repeatable

Within other elements: Optional, repeatable

Examples:
```
<conventiondeclaration>
    <abbr> ISAD(G) </abbr>
    <citation>ISAD(G): General International Standard
    Archival Description, second edition, Ottawa
    2000</citation>
</conventiondeclaration>
<didnote>File also contains materials from the
```

```
            <abbr expan="American Civil Liberties Union">
            ACLU </abbr>
        </didnote>
        <c02>
            <did>
                <unittitle>
                    <abbr expan="United Nations Educational,
                    Scientific and Cultural Organization"> UNESCO
                    </abbr>
                </unittitle>
            [. . .] </did>
        </c02>
```

<objectIdentifier> objectIdentifier

Summary:	A designation used to identify the Object uniquely within the preservation repository system in which it is stored.
Summary:	An identifier may be created by the repository system at the time of ingest, or it may be created or assigned outside of the repository and submitted with an object as metadata. Similarly, identifiers can be generated automatically or manually.
May contain:	objectIdentifierType, objectIdentifierValue
Data Type:	Container.
Description and Usage:	Description you can use <to_get_anglebrackets>
	Each Object held in the preservation repository must have a unique identifier to allow other entities to refer to it and to relate it to descriptive, technical, and other metadata unambiguously.
Description and Usage:	**Intellectual Entity / Representation**

Applicability Applicable Obligation Mandatory Repeatability Repeatable Example IE / Representation example

PREMIS File

Applicability Applicable Obligation Mandatory Repeatability Repeatable Example File example

PREMIS Bitstream

Applicability Applicable Obligation Mandatory Repeatability Repeatable Example Bitstream example

The objectIdentifier is mandatory for all Objects stored.

The objectIdentifier is repeatable in order to allow both repository assigned and externally-assigned identifiers to be recorded. See "Creation/Maintenance" note above.

Primary identifiers must be unique within the repository. They may be preexisting, and in use in other digital object management systems. Ideally, secondary identifiers should also be unique but sometimes this is not possible (e.g., if the values are inherited from a legacy system which did not enforce this or only identified items at a higher level). Identifiers for each item must be sufficient to identify the item uniquely at the appropriate level of aggregation. For example, an Intellectual Entity that represents all books

in the same edition could use an ISBN but this would be insufficient to identify a particular copy of that book.

A preservation repository needs to know both the type of object identifier and the value. If the value itself contains the identifier type (e.g., "oai:lib.uchicago.edu:1"), the identifier type does not need to be recorded explicitly. Similarly, if the repository uses only one type of identifier, the type can be assumed and does not need to be recorded.

Appendix :

Appendix A : Our appendix

A list of some sort

Label   Item