DSpaceMETSSIPProfile

Contents

DSpace METS Document Profile for Submission Information Packages (SIP)

Robert Wolfe, MIT Libraries
William Reilly, MIT Libraries

Acknowledgements

This document was prepared with the assistance of:

MacKenzie Smith, MIT Libraries
Rob Tansley, Hewlett-Packard
Larry Stone, MIT Libraries
Margret Branchovsky, MIT Libraries

Introduction

Reference Model for Open Archives Information Systems (OAIS)

DSpace at MIT has implemented the Reference Model for Open Archival Information Systems (OAIS) http://nssdcftp.gsfc.nasa.gov/standards/ccsds/pdf /CCSDS-650.0-B-1.pdf. DSpace's implementation has identified a need to prepare a METS profile or profiles that will govern the creation of the three types of content "packages" defined by the reference model.

This profile is intended to provide a complete set of instructions for the preparation of OAIS Submission Information Packages (SIPs). The purpose of this SIP profile is to make creation of SIPs as easy as possible for DSpace content partners. Many of these content partners will be other DSpace instances. In this case, the SIP profile may also serve as an OAIS Dissemination Information Package (DIP) profile. Future use of this profile, or related profiles, to govern the creation of Archive Information Packages (AIPs) will require the inclusion of additional information to account for the larger information needs of AIPs.

DSpace Content Object Model

In order the make easier the preparation of conformant SIPs, the DSpace Content Object Model is here mapped to the METS object model.

1. DSpace Item = METS Document (*Structural Requirements #19*)
   - This profile identifies a single METS document with a single Item in the DSpace Content Object Model. This profile does not anticipate or allow for the aggregation of multiple DSpace Items into one METS document for ingest.
2. DSpace Bundle = METS <fileGrp> (*Structural Requirements #12*)
   - DSpace groups files into "bundles," the purpose of which is to make it easier for DSpace to process them. These Bundles are loosely controlled by a vocabulary of bundles types. Creators of SIP documents who are familiar with DSpace bundle types are encouraged to organize <fileGrp> elements by bundle and to label them using the USE attribute. The USE attribute of the <fileGrp> element is reserved for use with the DSpace Bundle Type Vocabulary, but it is not required. DSpace will deposit the files in any unlabeled <fileGrp> element into the "Content" bundle.

3. Content Bundle = METS <structMap> (*Structural Requirements #17*)
   - The first <structMap> is reserved to organize the files intended for the DSpace "Content" Bundle. These files will be displayed to the public.
4. Bitstream = METS <file> element, <mdRef> element (*Structural Requirements #1*)
   - All files intended for inclusion in the DSpace Item as bitstreams should be found at either a xlink:href attribute of the <file> element, or a xlink:href attribute of the <mdRef> element. Metadata files referenced from <mdRef> elements in the <dmdSec> and <amdSec> elements will be deposited in the DSpace item as bitstreams. Only the metadata record for the whole item will be processed into the DSpace metadata database tables.
5. DSpace Item MD = METS <div> (*Structural Requirements #16*)
   - The <div> element is a child of the <structMap> element and is recursive. In the first <structMap> there must be only one first level <div> element. It may contain any number of child <div> elements. This first level <div> element represents the DSpace Item within the structMap and must contain references to metadata for the Item.
6. Primary Bistream = METS <fptr>
   - When DSpace ingests websites as items it identifies a "primary bitstream," or index file that it presents to the public. All other website files are suppressed to the DSpace user. The first <div> element under the first <structMap> element, also called the DSpace Item <div>, may only contain a child <fptr> element when the METS document represents a website with a primary bitstream. This <fptr> element should then reference the <file> element that represents the primary bitstream.
7. Preferred Bitstream = METS <file> element @USE (*Structural Requirements #*)
   - There are some potential DSpace Items that are not websites (they do not have a primary bitstream) that need to show preference for one bitstream or file above all others. One example of this is a potential item that contains three versions of the same document (e.g. pdf, ps and latex). It is advantageous to the user to know that the pdf is the version intended for public consumption. In this case this profile recommends that USE attribute of the <file> element representing the preferred bitstream contain the value "preferred".

Descriptive Metadata

As outlined in Rules of Description #1 and #2, the DSpace has adopted the Metadata Object Description Schema (MODS) as a transfer schema for descriptive metadata in SIPs.

As declared in Structural Requirements #16, DSpace requires just one MODS record that describes the entire item. This MODS record will be deposited into the DSpace Metadata Tables. This profile also recommends the inclusion of other metadata records where they exist. These records may describe discrete parts of the DSpace Item like single files, they may also record Item metadata in another schema native to system authoring the SIP. Future versions of this profile will address the use of the GROUPID attribute of the <dmdSec> element in coordination with the DMDID (IDREF) attribute to associate multiple records in different schemas. For now, any other metadata that is included in the SIP will not be processed into the metadata tables, but will be remain associated with the Item.

Technical Metadata

DSpace has defined a Technical Metadata Element Set to fulfill their preservation and content lifecycle management information needs. This Technical Metadata Element Set is best expressed using the PREMIS Preservation Metadata Schema: Object. DSpace use of the PREMIS Data Dictionary to represent needed technical metadata elements does not constitute full implementation of the PREMIS data model.

There is also a Structural Requirement that encourages the use of CHECKSUM, CHECKSUMTYPE, CREATED and MIMETYPE attributes of the <file> element. These attributes mirror elements in the Technical Metadata Element Set. In order to avoid confusion on the part of content partners preparing SIPs, the following logic applies to the addition of technical metadata to METS instances conformant with this SIP profile.

- Do you as a content provider have this information?
- If you do not have it, DSpace will create some of this information upon ingestion of the package.
- If you do have this information, its inclusion is still optional but strongly recommended.
- If you can, write both.
- If you can't write both, write the <techMD> using PREMIS elements.
- If you can't write that, write the attributes in the <file> element.

DSpace Bitstream Metadata

In the use case where one DSpace instance uses this profile to create a SIP intended for a second DSpace instance it would be useful to include metadata the DSpace captures for each bitstream. There are three semantic units that DSpace captures: name, source and description. The appropriate metadata schema for transferring this information is currently under investigation.

In addition DSpace assigns a Sequence ID to each bitstream. These sequence IDs may look like handles, but they are not handles and will not resolve via the Handle system. The appropriate means to include Sequence IDs in bitstream metadata is under investigation for the DSpace-2-DSpace use case. If Sequence IDs are included in a SIP, every bitstream would require a unique sequence ID to avoid collision with a DSpace import mechanism that will assign a Sequence ID to any bitstream lacking one.

Rights Metadata

The DSpace Item Submission interaction provides an opportunity to assign a Creative Commons license to the material deposited in the repository. In this METS SIP profile the same opportunity is provided. Inclusion of CC licenses as rdf/xml is encouraged, but not required, in Rules of Description #11. An example of CC license metadata is included in the sample METS document in the Official XML Expression of the METS SIP Profile.

The DSpace Deposit License is not required for METS documents that conform to this SIP Profile. It is assumed that agreement concerning this license between DSpace and its content providers will be accomplished elsewhere than the submission package.

Official METS Profile Documentation

The following are the necessary component parts of any METS profile conforming to the METS Profile Schema as defined at http://www.loc.gov/standards /mets/profile_docs/components.html

These parts are presented first in human readable form, then repeated in an requirements compliant xml expression. The xml expression is governed by the schema at: http://www.loc.gov/standards/mets/profile_docs/mets.profile.v1-2.xsd

URI

Title

DSpace METS Document Profile for Submission Information Packages (SIP)

Abstract

This profile specifies how METS documents organizing Items for submission to DSpace should be encoded.

Creation Date

16 April 2007, 15:34:00 EDT

Contact Information

Robert Wolfe
DSpace Federation
77 Massachusetts Ave, Cambridge, MA 02138
(617) 253-0604
rwolfe@mit.edu

Related Profile

Extension Schema

1. MODS (Metadata Object Description Schema), version 3http://www.loc.gov/standards/mods/v3/mods-3-1.xsd
   - Elements from MODS are used to express descriptive metadata for the DSpace Item and its constituent files.
2. Creative Commons License Schemahttp://web.resource.org/cc/schema.rdf
   - The Creative Commons RDFXML schema is used to capture distribution licenses for the content of METS SIPs.
     #PREMIS Preservation Metadata Schemahttp://www.loc.gov/standards/premis/PREMIS-v1-0.xsd
   - Elements from the PREMIS Preservation Metadata Schema: Object are used to express the DSpace Required Technical Metadata Element Set.

Rules of Description

1. The official descriptive metadata record for the DSpace Item must conform to the Metadata Object Description Schema (MODS) defined at http://www.loc.gov/standards/mods/v3/mods-3-2.xsd .
2. Other metadata records included in conforming METS documents may be expressed in any metadata schemes.
3. Technical metadata for DSpace Items and Bitstreams occurring in the techMD element should conform to the DSpace Required Technical Metadata Element Set defined at http://wiki.dspace.org/index.php/DSpaceMETSSIPProfile. This metadata should be encoded using the PREMIS Data Dictionary.

Controlled Vocabularies

- **Name:** DSpace Bundle Types
- **Maintenance Agency:** DSpace Federation
- **URI:** http://www.dspace.org/standards/vocab/bundles/bundleTypev1p0/DSpacebundletypev1p0.pdf
- **Values:**
  - ORIGINAL
  - TEXT(EXTRACTED)
  - THUMBNAIL
  - LICENSE
  - CC_LICENSE
  - METADATA

Structural Requirements

1. A conforming METS document must represent only one DSpace Item.
2. A conforming METS document is a complete manifest of the DSpace Item. Do not include content or metadata files in the SIP that are not referenced in the METS document.
3. The DSpace Content Object Model organizes DSpace Items into Bundles. Bundles are exclusive classifications of files within a DSpace Item.
4. Content files--files intended for the Content bundle--must be included in or referenced from the fileSec and the structMap. Metadata files--files intended for the Thumbnail, Text (Extracted), License, CC_License and Metadata bundles--are included in or referenced either from the fileSec, dmdSecs or amdSecs.
5. DSpace has defined a set of technical metadata elements for preservation and administration. This metadata includes a unique identifier, checksum, checksum type, mimetype, file size, creation date and file path originally assigned to the file. If this data exists within a system that is authoring a conforming METS document it should be included within the METS document.

6. Inclusion of technical metadata should occur in the techMD element and should conform to the DSpace Required Technical Metadata Element Set defined in this document. This metadata should be encoded using the PREMIS Data Dictionary.
7. A conforming document may contain user supplied Creative Commons licenses in the rightsMD element.
8. A conforming METS document must reference all files accompanying the METS document and comprising the DSpace Item via an xlink:href attribute on either an mdRef or FLocat element. There must be only one FLocat element per parent File element.
9. A conforming METS document must contain the ID attribute of the METS root element.
10. A conforming METS document should contain the PROFILE attribute of the METS root element.
11. DSpace implementations will ignore the metsHdr element, its attributes, child elements and their attributes.
12. The dmdSec is reserved exclusively for bibliographic description and subject analysis of the item and its constituent files, at a ratio of one dmdSec for each metadata record. Multiple expressions of the same metadata in multiple schemas must be recorded in separate dmdSecs and must be grouped through the GROUPID attribute.
13. A conforming METS document must contain at least one dmdSec containing the metadata record for the entire DSpace item the document represents.
14. Each unique configuration of techMD and rightsMD elements must be contained within a separate amdSec element.
15. A conforming METS document must contain the ID attribute for all amdSec elements.
16. DSpace implementations will ignore the sourceMD element, its attributes, child elements and their attributes.
17. DSpace implementations will ignore the digiprovMD element, its attributes, child elements and their attributes.
18. File elements must not contain the FContent child element. A conforming METS document may not contain content encoded as binary or xml data. These encoding mechanisms may be used to include metadata in the METS document.
19. It is strongly recommended that the USE attribute be present for every fileGrp element included in conforming METS documents. The USE attribute identifies Bundles within the METS SIP. Eligible values for this attribute are restricted to the DSpace Bundle Type vocabulary.
20. Multiple expressions of the same content object (e.g. thumbnails and archival masters of the same image) though organized in separate DSpace bundles should be related via the GROUPID attribute of the File element.
21. In the case of multiple expressions of the same content object in different file formats (e.g. .pdf, .ps, .latex), it is strongly recommended that the USE attribute be present on the one file element representing the format that is preferred for public consumption. The value of this attribute must be "preferred".
22. If available, supply the CHECKSUM, CHECKSUMTYPE, CREATED and MIMETYPE attributes of the File element.
23. The first div element under the first structMap element shall be used to identify the DSpace Item and, for websites, the primary bitstream. It must not contain an fptr element unless the DSpace Item is a website with a primary bitstream. It must contain AMDID and DMDID (IDREF) attributes that identify the appropriate metadata for the Item to be processed into the DSpace metadata database tables upon ingest.
24. All files in the content bundle must be represented by child div elements of the first div (DSpace Item div) element of the first structMap element.
25. Multiple structMap elements recording alternate organizations of the DSpace Item are encouraged when applicable.
26. A conforming METS document represents a single DSpace Item and must not contain any mptr elements referencing other METS documents.
27. DSpace implementations will ignore the structLink element, its attributes, child elements and their attributes.
28. DSpace implementations will ignore the behaviorSec element, its attributes, child elements and their attributes.

Technical Requirements of Content, Behavior and Metadata Files

1. The list of allowable content files that may be referenced in conforming documents via the FLocat element is restricted to those files each DSpace instance has agreed to support.
2. Metadata Files should be encoded in xml and should validate to the schema corresponding to the mdType attribute value of the mdRef element.

Tools and Applications

1. This profile is intended for use with the DSpace Packager Plugin, sometimes called the Lightweight Network Interface and discussed at Packager Plugins.

Examples

CSAIL Example

DSpace to DSpace Example