



Harvard University Library

RLG Forum/ALA, June 16, 2002, Atlanta

An Archival Submission
Information Package for E-Journals

Stephen L. Abrams
Harvard University Library
stephen_abrams@harvard.edu



Harvard University Library

E-Journal Archive

- Mellon E-journal archive project
 - <http://www.diglib.org/preserve/ejp.htm>
- “Preserve significant intellectual content ... independent of the form originally delivered”
- Publisher, not subject based
- Multiple external content suppliers
- Initially dark content



Design Principles

- **Issue-centric**
- **Capture content at highest resolution, finest granularity, most abstract representation**
 - Archiving work, not manifestation
- **Architecture based on OAIS**
- **Acceptance by stakeholders is cost-sensitive**
 - Automation
 - Standards
 - Homogeneity

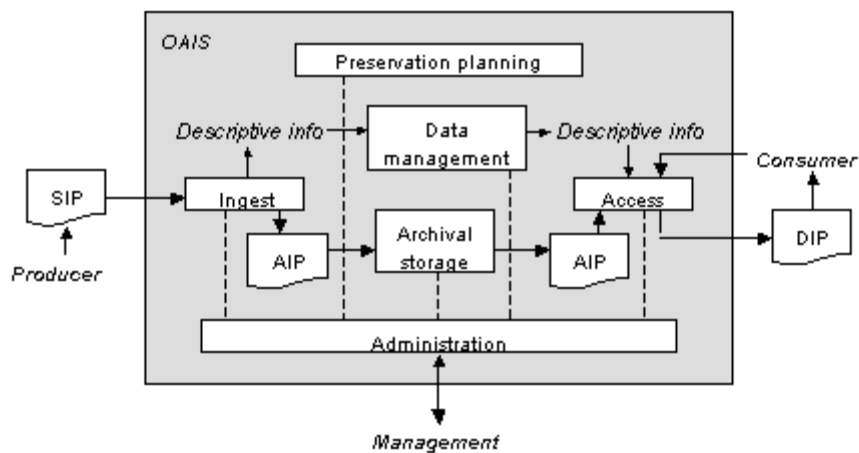


What Is OAIS?

- Open Archival Information System
 - <http://ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf>
- “A common framework of terms and concepts ... to provide long-term preservation of digital information”
- An functional and information reference model



OAIS Functional Model





Harvard University Library

Archival SIP

- Unit of submission is the e-journal issue
- Modeled at two levels: issue and item
- Explicit separation of content and metadata
- METS used for metadata
- XML DTD for item content under development by Harvard and NLM, based on PMC2 DTD



Normative Data Formats

- **Necessary for internal data homogeneity**
- **Single format for each content category**
 - Text is XML, raster still image is TIFF, ...
- **Standards, maturity, viability, robust tools, created upstream in production process**
- **Lower level specification than MIME type**
 - Bi-tonal TIFF with Group IV compression
- **Non-normative formats transformed on ingest**



Format Registry

- **Version history**
- **Authoritative specification / maintenance org.**
- **Identity characterization**
 - MIME type, magic number, internal syntax
 - Application specific profile
- **Technical metadata schema**
- **Compliant tools**
- **Community-wide resource and responsibility**



SIP Directory Structure

```
titleid/  
  issueid/  
    issue-md.xml      metadata (METS)  
    issue.xml  
    cover.tif        content  
    ...  
    itemid1/  
      item-md.xml    metadata (METS)  
      item.xml  
      item.pdf  
      fig1.tif      content  
      ...  
    itemidn/  
      ...
```



Harvard University Library

What Is METS?

- **Metadata Encoding & Transmission Standard**
 - <http://www.loc.gov/standards/mets/>
- “A standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library”
- DLF funded; maintained at LC MARC Standards Office



METS Schema

- **Namespace-qualified XML schema**
 - <http://www.loc.gov/standards/mets/mets.xsd>
- **Explicit structural metadata; containers for externally-defined descriptive and administrative metadata (“extension schemas”)**
- **External pointers using XLink; internal links via ID/IDREF**



Harvard University Library

Why METS?

- Why not use RDF, Topic Map (ISO/IEC 13250), or a custom schema?
- METS is designed specifically for library-like digital objects
- Appropriate technology with pre-defined semantics
- Community support

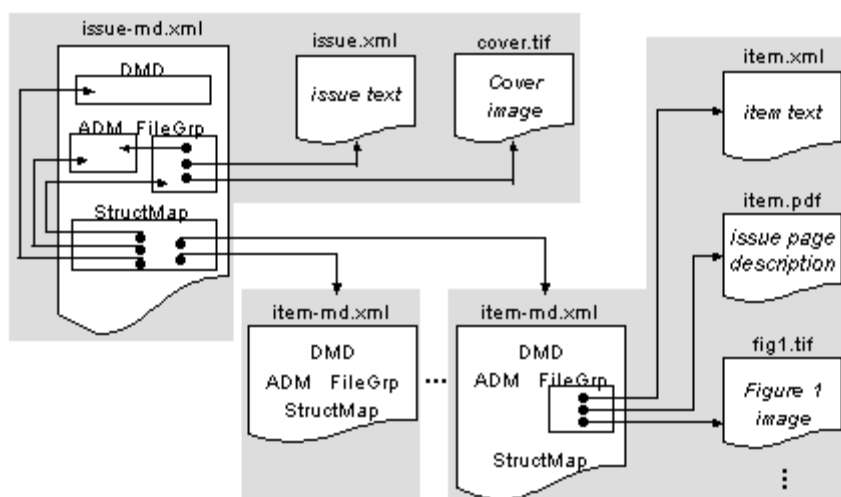


METS Structure

```
<mets ID="..." OBJID="..." LABEL="..." TYPE="..." PROFILE="...">
  <metsHdr> ... </metsHdr>
  <dmdSec> ... </dmdSec>
  <amdSec>
    <techMD> ... </techMD>
    <rightsMD> ... </rightsMD>
    <sourceMD> ... </sourceMD>
    <digiprovMD> ... </digiprovMD>
  </amdSec>
  <fileSec> ... </fileSec>
  <structMap> ... </structMap>
  <behaviorSec> ... </behaviorSec>
</mets>
```



SIP Hierarchical Structure





SIP Details

- **File and internal ID naming conventions designed to make SIP self-documenting**
- **Non-ASCII Unicode entered as XML numeric entities; non-Unicode as character entities**
- **SIP is aggregated and compressed into a JAR file for submission**
- **Full SIP specification (Version 1.0 DRAFT)**
 - <http://www.diglib.org/preserve/harvardsip10.pdf>



SIP Details

- **File and internal ID naming conventions designed to make SIP self-documenting**
- **Non-ASCII Unicode entered as XML numeric entities; non-Unicode as character entities**
- **SIP is aggregated and compressed into a JAR file for submission**
- **Full SIP specification (Version 1.0 DRAFT)**
 - <http://www.diglib.org/preserve/harvardsip10.pdf>



Harvard University Library

METS Java Toolkit

- **Procedural construction and parsing of METS files**
 - <http://hul.harvard.edu/mets/>
- **Java API**
- **Local and global validation**
- **Marshal/unmarshal**
 - Serialize in-memory representation to file
 - De-serialize from file to in-memory representation



Toolkit Implementation

- **Generic API**
 - Can be sub-classed for application-specific behavior
- **Based on Sun's JAXB specification**
 - <http://java.sun.com/xml/jaxb/>
- **Uses Jim Clark's XP parser**
 - <http://jclark.com/xml/xp/>



Procedural Construction

```
Mets mets = new Mets();
mets.setOBJID("200206.6");
mets.setType("Issue");
    MetsHdr metsHdr = new MetsHdr();
        metsHdr.setCREATEDATE(new Date());
        metsHdrs.setRECORDSTATUS("SIP");
    ...
mets.getContent().add(metsHdr);
DmdSec dmdSec = new DmdSec();
...
mets.getContent().add(dmdSec);
...
mets.validate();
mets.marshal(new FileOutputStream("issue-md.xml"));
```



METS Automation Tools

- **For depositors:**
 - Construction of partial METS files
 - Still need to add specific metadata
 - Pre-deposit validation
 - API integration into existing content management and production systems
- **For the archive:**
 - Parsing and validation during ingest
 - Ingest conversion of SIP to AIP
 - Access conversion of AIP to DIP



Are These Standards Helpful?

- **OAIS**
 - Common vocabulary for inter-project and inter-disciplinary conversation
 - Conceptual mapping between heterogeneous systems
- **METS**
 - Appropriate technology
 - Not a universal panacea; but sufficient for library-like resources and processes
 - Community support



Harvard University Library

Questions?

- Mellon E-Journal Archiving Project
 - <http://www.diglib.org/preserve/ejp.htm>
 - <http://www.diglib.org/preserve/harvardsip10.pdf>
- OAIS
 - <http://ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf>
- METS
 - <http://www.loc.gov/standards/mets/>
 - <http://hul.harvard.edu/mets/>
- stephen_abrams@harvard.edu