Building a Digital Repository on a Shoestring Budget

Christinger Tomer
University of Pittsburgh

PALA
September 30, 2014

A version this presentation is available at http://www.pitt.edu/~ctomer/shoestring/
Repositories serve various purposes, but in the main they are built to guarantee the availability of and access to select materials. And, like libraries and archives, a more specific purpose is derived from the community that is served.

Repository
 Reasons for Building a Digital
Types of Digital Repositories

- Digital Archives
  - compliant with archival standards and practices
- Digital Repositories
  - compliant with library standards for metadata
  - generally outside the realm of standards for library and archival practice; may comply
- Asset Management Systems
  - outside the realm of standards for library and archival practice

Types of Digital Repositories

- Digital Archives
- Digital Repositories
- Asset Management Systems
What is the goal of digital preservation? The ultimate outcome of the preservation process is, the outputs of a preservation process ought to be identical, in all essential respects, to what went into that process; that is, the outputs of a preserved object should be authentic preserved objects; that is, the outputs of a preserved object should be identical, in all essential respects, to what went into that process.
A trusted digital repository is one whose mission is to provide reliable, long-term access to managed digital resources on behalf of its depositors and users. It must meet expectations for system evaluation and security of materials deposited within it. The overall system, however, to meet expectations all trusted digital repositories must:

- Have organizational systems that support not only long-term viability of the repository, but also the digital information for which it has responsibility;
- Establish methods for system evaluation that meet community expectations of trustworthiness;
- Demonstrate fiscal responsibility and sustainability;
- Design its system(s) in accordance with commonly accepted conventions and standards to ensure the ongoing management, access, and security of materials deposited and users;
- Accept responsibility for the long-term maintenance of digital resources on behalf of its depositors and for the benefit of current and future users;
- Establish methods for system evaluation that meet community expectations of trustworthiness;
- Be depended upon to carry out its long-term responsibilities to depositors and users; and
- Have policies, practices, and performance that can be audited and measured.

The Trustworthy Repository
Attributes of a Trusted Repository

- Procedural accountability
- System security
- Technological and procedural suitability
- Financial sustainability
- Organizational viability
- Administrative responsibility

Compliance with the Reference Model for an Open Archival Information System (OAIS)
Compliance with the Reference Model for an Open Archival Information System (OAIS)

The OAIS Reference Model supplies a common framework, including terminology and concepts, for describing and comparing architectures and operations of digital archives. It also provides both a functional model—the specific tasks performed by the repository such as storage or access—and a corresponding information model that includes a model for the creation of metadata to support long-term maintenance and access.
Responsibilities of a Trusted Repository

Research repositories need to understand fully what responsibilities they should assume for the preservation of digital materials. Organizational responsibility must be understood at three basic levels: understanding local requirements and how to meet them; understanding which responsibilities can be shared and who share them; and understanding which responsibilities can be shared and how they can be shared.

In summary, the major factors in defining the responsibilities of a trusted repository are:

- the scope of collections;
- preservation and lifecycle management;
- the wide range of stakeholders;
- ownership of material and other legal issues; and
- cost implications

Organizations might share responsibilities through geo-replication or consortial agreements or shared user communities, or local requirements and how to meet them; understanding how other responsibilities must be understood at these basic levels. Organizations should assume for the preservation of digital materials. Organizational responsibilities need to understand fully what responsibilities they should assume for the preservation of digital materials.
A Reliable Digital Repository negotiates for and accepts appropriate information from information producers and rights holders; obtains sufficient control of the information provided to support long-term preservation; determines, either by itself or with others, the users that make up its designated community; ensures that the information to be preserved is "independently understandable" to the designated community, which should be able to understand the information provided; follows documented policies and procedures that ensure the information is preserved against all reasonable contingencies and enables the information to be disseminated as authenticated copies of the original or as accessible to the original; makes the preserved information available to the designated community; and works closely with the repository's designated community to advocate the use of standard practice in the creation of digital resources, including outreach programs.
OAI Protocol for Metadata Harvesting

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a protocol developed by the Open Archives Initiative (OAI). It is a protocol developed by the OAI for collecting metadata from many archives. An implementation of OAI-PMH must support providing metadata in Dublin Core, but may also support additional representations. OAI-PMH uses XML over HTTP. Version 2.0 was updated in 2008.
Open source software is made available under a license — there are many variations — that allows the software to be used free of charge and makes its source code available for modification and improvement, provided that the improvements revert to the developer community. Open source systems predominate in this area of computing, largely because commercial developers have not viewed it as sufficiently lucrative.
Key Factors in Adoption and Use

- Presentation Modes
- Extensibility
- Import/Export Capabilities
- Workflows
- Compliance with Technical Standards,
  including OAI-PMH and OAIS
Cost Factors

- Labor — setup/startup time is relatively short, but not always sufficient; open source options are generally available, but not always. For collection, conversion, etc., of such files will be needed; building is a labor-intensive process.
- Computer — a small-to-medium-sized Intel-based server is most appropriate.
- Disk Storage & Backup — storage costs are low and wholly redundant. Continue to drop, but the storage system should be server is most appropriate.
- Other Software — for collections that entail audio, video, and/or image files, utilities supporting the processing, conversion, etc., of such files will be needed.
- Collection building is a labor-intensive process.
Costs of External Support

- DuraCloud — basic plan is $1800 a year, with 1TB storage
- omeka.net — plans range from $50-999 a year, with 1-25GB storage and support for unlimited sites, plugins, and themes on the high end
- ePrints Services — prices on request

With 1TB storage

Costs of External Support
Key Open Source Systems
The digital archive is distinguished by the use of metadata schemes that have been developed within the archival community and by an emphasis on the collection, as opposed to discrete items, as the primary structure for organization and presentation. The digital archive is

Digital Archive
Digital asset management systems tend to place less emphasis on the support for formally defined metadata schemes, often relying instead on more ad hoc approaches to description, and more emphasis on the presentation of content.
Digital Repositories are systems designed to store and provide access to digital objects of various types, and they are distinguished by the use of standardized workflows. Of the current systems, ePrints is the most popular, because it supports self-archiving. DSpace is also popular. Its customizable workflow, increasing extensibility, and support for standardized formatting and packaging are main reasons.

Of the current systems, DSpace is also popular. Its customizable workflow, increasing extensibility, and support for standardized formatting and packaging are main reasons. Of the current systems, DSpace is also popular. Its customizable workflow, increasing extensibility, and support for standardized formatting and packaging are main reasons.

Of the current systems, DSpace is also popular. Its customizable workflow, increasing extensibility, and support for standardized formatting and packaging are main reasons. Of the current systems, DSpace is also popular. Its customizable workflow, increasing extensibility, and support for standardized formatting and packaging are main reasons.

Of the current systems, DSpace is also popular. Its customizable workflow, increasing extensibility, and support for standardized formatting and packaging are main reasons. Of the current systems, DSpace is also popular. Its customizable workflow, increasing extensibility, and support for standardized formatting and packaging are main reasons.

Of the current systems, DSpace is also popular. Its customizable workflow, increasing extensibility, and support for standardized formatting and packaging are main reasons. Of the current systems, DSpace is also popular. Its customizable workflow, increasing extensibility, and support for standardized formatting and packaging are main reasons.
The LAMP model is the most commonly used configuration, owing to the ease with which the underlying database and the web server may be integrated.
Like LAMP configurations, Apache Tomcat may be configured to work with open source RDBMS such as MySQL and PostgreSQL. Under heavier loads, Tomcat configurations are thought to be more stable.
ATOM -- Access to Memory

This version offers new presentation capabilities.

This is a top-level view of a version of ATOM that is being developed in conjunction with Archives Canada.
Basic Entry under AToM

This is the basic entry form for AToM, supplemented by forms dealing with administrative and IP issues as well as provenance and archival description.
Support for Archival Standards

metadata and other forms of archival standards
ATOM provides support for
Template for AtoM's Administrative Area

| Process notes |
| Process group |
| Process status |

Includes information about type, creators, custodians, and priority, etc. Condition, processing status, appraisal, physical history, appraisal, physical type, creators, custodians.
Archival standards. Records have been designed to conform to aspects of ATOM, the creation of authority records under ATOM, like other Corporate Bodies, Persons and Families.

ISOAR is the ISOAR (CPF): International Standard Archival Authority Record for Places.

This is a fragment of the form for creating authority records.
DSpace is a digital repository system that began as a collaboration between MIT and Hewlett-Packard in 2002 and is now maintained by DuraSpace, which is a community of institutions dedicated to supporting and advancing open-source digital repository software. DSpace supports key standards for metadata, packaging, and import/export.
DSpace is a set of cooperating Java web applications and utility programs that maintain an asset store and an associated metadata store. The Web applications provide interfaces for administration, deposit, ingest, search and access. The asset store is maintained on a file system or similar storage system, while the metadata, including access and configuration information, is stored in a relational database and supports the use of PostgreSQL and Oracle databases. DSpace also supports the OAI-PMH v2.0, and is capable of exporting METS (Metadata Encoding and Transmission Standard) packages. DSpace provides the common interoperability standards used in the Institutional repository domain, such as Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and the Java Servlet API and XML (aka Manakin), JSP which uses JSP and the Java Servlet API and XSLT (aka Manakin), JSPUI which uses JSP and the Java Servlet API and XSLT (aka Manakin). DSpace currently supports two primary web interfaces:

- **JSPUI** which uses JSP and the Java Servlet API and XML (aka Manakin),
- **XSLT** which is based on Apache Cocoon, using XSLT and XSLT.

DSpace holdings are made available primarily via a web interface, but it also supports the OAI-PMH v2.0, and is capable of exporting METS (Metadata Encoding and Transmission Standard) packages. DSpace supports the common interoperability standards used in the Institutional repository domain, such as Open Archives Initiative Protocol for Metadata Harvesting, SWORD, OpenSearch, and RSS. More recent versions of DSpace also support faceted search and browse functionality using Apache Solr.
The DSpace interface can be enhanced; the most popular enhancement, also known as “Manakin,” was developed at Texas A&M.
Editing a DSpace Item
Metadata under DSpace
<table>
<thead>
<tr>
<th>Group</th>
<th>Action</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anonymous [EdX]</td>
<td>25735</td>
<td>READ</td>
</tr>
<tr>
<td>Anonymous [EdX]</td>
<td>25737</td>
<td>DEFAULT ITEM READ</td>
</tr>
<tr>
<td>Anonymous [EdX]</td>
<td>25738</td>
<td>DEFAULT BROWSER READ</td>
</tr>
<tr>
<td>collection [3]: ADMIN [EdX]</td>
<td>25739</td>
<td>ADMIN</td>
</tr>
<tr>
<td>collection [3]: submit [EdX]</td>
<td>25740</td>
<td>ADP</td>
</tr>
<tr>
<td>collection [3]: workflow_step [EdX]</td>
<td>25741</td>
<td>ADP</td>
</tr>
<tr>
<td>collection [3]: workflow_step [EdX]</td>
<td>25742</td>
<td>ADP</td>
</tr>
<tr>
<td>collection [3]: workflow_step [EdX]</td>
<td>25743</td>
<td>ADP</td>
</tr>
</tbody>
</table>

Click here to add a new policy.

**Default Policies Set**
<table>
<thead>
<tr>
<th>Metadata Fields</th>
<th>Scope Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date of creation or manifestation of intellectual content</td>
<td>67</td>
</tr>
<tr>
<td>Date of copyright</td>
<td>16</td>
</tr>
<tr>
<td>Date of publication or distribution</td>
<td>15</td>
</tr>
<tr>
<td>Date or date range item becomes available to the public</td>
<td>14</td>
</tr>
<tr>
<td>Date or date range item becomes available to the public</td>
<td>13</td>
</tr>
<tr>
<td>Date or date range item becomes available to the public</td>
<td>12</td>
</tr>
<tr>
<td>Date of item</td>
<td>11</td>
</tr>
<tr>
<td>Date of item</td>
<td>10</td>
</tr>
<tr>
<td>Date of item</td>
<td>9</td>
</tr>
<tr>
<td>Date of item</td>
<td>8</td>
</tr>
<tr>
<td>Date of item</td>
<td>7</td>
</tr>
<tr>
<td>Person, organization, or service responsible for the content of the resource</td>
<td>6</td>
</tr>
<tr>
<td>Person, organization, or service responsible for the content of the resource</td>
<td>5</td>
</tr>
<tr>
<td>Person, organization, or service responsible for the content of the resource</td>
<td>4</td>
</tr>
<tr>
<td>Person, organization, or service responsible for the content of the resource</td>
<td>3</td>
</tr>
<tr>
<td>Person, organization, or service responsible for the content of the resource</td>
<td>2</td>
</tr>
<tr>
<td>Person, organization, or service responsible for the content of the resource</td>
<td>1</td>
</tr>
</tbody>
</table>

**Schema Metadata Fields**
Harvester Settings

- OAI source: all

Generator Settings

- Internal errors: OAI
- OAI errors: OAI

Queued Harvests:

- OAI Harvester:
- Active Harvester:

Collections set up for harvesting:

- Smart Harvester
- Host Harvester

Actions:

- OAI Harvester Status: Automatic harvesting is not active (inactive)
- Harvester Schedule Controls

Control Panel

Current Activity | Harvester Status | System-wide Alerts | OAI | Dispace Configuration | Java Information

OAI Harvesting
Omeka mimics a development model that was devised by a number of open source content management systems, most notably Drupal, embracing a more modular approach to design and development. Omeka surrounds a core set of functions which may be deployed on a collection-by-collection basis and used at the discretion of the system’s administrators, and some of which may be deployed on a collection-by-collection basis.

Omeka mimics a development model that was devised by a number of open source content management systems, most notably Drupal, embracing a more modular approach to design and development.
In this example, Omeka presents a text of the paper linked to the full automatically generated facsimile of a PDF version of a paper, with an entry for a PDF.
Omeka’s LC Suggest Module

Omeka takes advantage of an API published by The Library of Congress to perform “look up” of matching subject headings. This is accomplished via a plugin module called “LC Suggest” and helps standardize this aspect of metadata generation.
EPrints is a Web and command-line application based on the LAMP architecture (but is written in Perl rather than PHP). It has been successfully run under Linux, Solaris and Mac OS X. A version for Microsoft Windows was released 17 May 2010. A version for Mac OS X, A version for Linux, Solaris and Mac OS X. A version for Windows has been released.

Version 3 of the software introduced a (Perl-based) plugin architecture for importing and exporting data, converting objects (for search engine indexing) and user interface widgets. Configuring an EPrints repository involves modifying configuration files written in Perl or XML. The appearance of a repository is controlled by HTML templates, CSS stylesheets and inline images. While EPrints is shipped with an English translation, it has been translated into other languages through redistributable (re)translatable language phrase files. Translations include Bulgarian, French, German, Hungarian, Italian, Japanese, Russian, Spanish and Ukrainian.
Key Proprietary Systems
Key Resources & Sites

- AtoM — https://www.accesstomemory.org/
- CONTENTdm
- DSpace — https://www.dspace.org/
- ePrints — https://www.eprints.org/software/
- Omeka — https://omeka.org/
Any questions?