METS-Based Cataloging Toolkit for Digital Library Management System

Li Dong, Bei Zhang
Library of Tsinghua University, Beijing, China
{dongli, zhangbei}@lib.tsinghua.edu.cn
Chunxiao Xing, Lizhu Zhou
Computer Science and Technology Department, Tsinghua University, Beijing, China
{xingcx, deszlz}@tsinghua.edu.cn

Abstract This toolkit is designed for the Digital Library Management System of Tsinghua University (TH-DLMS). The aim of TH-DLMS is to build up a platform to preserve various kinds of digitalized resources, manage distributed repositories and provide kinds of service for research and education. This toolkit fulfills the cataloging and preservation functions of TH-DLMS. METS (Metadata Encoding and Transmission Standard)[1] encoded documents are used as the final storage format of metadata, including descriptive metadata, structural metadata and administrative metadata, and submitted to a management system based on Fedora (Flexible Extensible Digital Object and Repository Architecture) Open Toolkit for TH-DLMS. This full-featured toolkit upon which interoperable web-based digital libraries can be built. The management functions of digital objects are powerful enough for us, but we have to solve the problem of creating the digital objects, since Fedora system doesn’t provide a tool for cataloging and preservation. In the following sections of this paper we’ll mainly describe our work on design and implementation of the cataloging and preservation toolkit for TH-DLMS.

1 Introduction

In recent years, long-term preservation of digital resources began to gain focus from libraries widely over the world, some related projects have been or being carried out for digital resources preservation and management in recent years, such as MOA2 (Making of American 2)[4] in UCB, the Library of Congress Audio-Visual Prototyping Project[5], the Fedora (Flexible Extensible Digital Object and Repository Architecture) project[2] funded by the Andrew W. Mellon Foundation. Enough have been done for preservation of simple-structured images and texts with standard ASCII characters, while few have been researched on complex-structured digital objects and texts with nonstandard ASCII characters. To solve these issues, TH-DLMS is designed to build up a platform to preserve various kinds of digitalized resources, manage distributed repositories and provide kinds of service for research and education. Fedora, an open-source project, provides flexible interfaces for us to do extensions on it for TH-DLMS, it’s a tool to act as a foundation upon which interoperable web-based digital libraries can be built. The management functions of digital objects are powerful enough for us, but we have to solve the problem of creating the digital objects, since Fedora system doesn’t provide a tool for cataloging and preservation. In the following sections of this paper, we’ll mainly describe our work on design and implementation of the cataloging and preservation toolkit for TH-DLMS.

2 Analyses and Design

2.1 About Metadata

For preservation purpose, not only the digitalized resources themselves should be saved, the metadata of the digitalized resources should also be saved and packaged with them properly so that those resources can keep original status after transmission and refreshment. There are 3 types of metadata for preservation, descriptive metadata, structural metadata and administrative metadata.

(a) Descriptive Metadata: Descriptive metadata is for digital object description to find or identify the resource. Since the purpose of TH-DLMS is to manage various kinds of digital resources, the
digital resources, the descriptive metadata should be common enough. We use Dublin Core (DC) metadata element set[3] as the standard for TH-DLMS resource description for it is widely used for cross-domain information resource description. As the base module of TH-DLMS, Fedora also uses DC as the descriptive metadata standard, which can be used to implement Open Archives Initiative (OAI)[6] Provider.The XML Schema definition of OAI-DC can be accessed at Open Archives Initiative (http://www.openarchives.org/OAI/2.0/oai_dc.xsd).

(2) Structural Metadata: Structural metadata is for representing the relationships inside a digital object, such as chapters of a book. In TH-DLMS, we use the <structMap> section of a METS[1] object to record the structural metadata; each <div> represents a node of digital contents.

(3) Administrative Metadata: Administrative metadata is data that supports the unique identification, maintenance, and archiving of digital objects, as well as related functions of the organization managing the repository. There are 4 parts of administrative metadata: technical metadata, rights metadata, source metadata and digital provenance metadata. Because the types of digital resources are rather different, such as plain text, formatted text, still images, audio and video; the technical metadata standards are rather difficult to choose, the technical metadata design is still in progress. Currently we use some simplified administrative metadata standards recommended by METS Official Web Site[1] and make some localization definitions for them at our technique web site[7][8][9][10] by giving each element a Chinese language name.

2.3 Characters and Encoding

2.3.1 Characters and Encoding

2.3.1.1 Encoding Metadata

Maintaining a library of digital objects of necessity requires maintaining metadata about those objects. We use METS to incorporate metadata of digital objects in TH-DLMS. The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation[1]. Many institutions and colleges have taken or started to take METS as a basic metadata packaging standard in their digital resources preservation systems[11], because it provides an XML document format for encoding metadata necessary for both management of digital library objects within a repository and exchange of such objects between repositories. Depending on its use, a METS document could be used in the role of Submission Information Package (SIP), Archival Information Package (AIP), or Dissemination Information Package (DIP) within the Open Archives Initiative Information System (OAI-PS) Reference Model[12].

2.3.1.2 Document representation and storage

The advanced architecture and the multiple scripts of Unicode provides the best opportunity to e-Publishers in the world, also provides the best opportunity for digital resources preservation. Considering data exchange between different systems from different countries, Unicode is the best choice of character encoding method compared with some other local standards for Chinese characters, such as GB2312, GBK, GB18030, because it’s global interoperable and supported widely by most popular software platforms. Since the basic document storage format of TH-DLMS is XML, whose legal accepted character set is specified by ISO/IEC 10646:2000 (Unicode).
Unicode in TH-DLMS is a wise choice for software implementation and data maintenance, as a result, UTF-8 and UTF-16 are used for character encodings.

3 Functions of the Toolkit

This cataloging and preservation toolkit provide main functions as following:

3.1 Metadata Cataloging

You can use this toolkit to input and edit metadata, including descriptive metadata and administrative metadata following TH-DLMS metadata standards mentioned in section 2.1. For most administrative metadata items, such as those for digital provenance metadata, technical metadata and rights metadata, which are almost the same for a batch of digital resources, we can import those metadata from a template file instead of input them one by one. But for descriptive metadata, a more common way is to input each element one by one.

3.2 Cataloging the Structure

To describe and record a resource’s inner structure, this toolkit provides the function of cataloging a resource’s inner structure. You can use this toolkit to create a complete structure of the resource, for example, from the whole book node, chapter node, and section node to page node. The node can extend to any level with free types, which are not limited by pre-defined structure types; and each node can be added, edited, or deleted freely on the structure tree, as the cataloger needs. For cataloger’s convenience, the toolkit provides auto-construct of page nodes, so that the cataloger need not create so many page nodes as they construct the resource’s structure, thus to save much time of cataloging the structure.Figure 1 is the structure representation of a book’s content.

3.3 Resource Linking

To organize the metadata and digital resources with correct structure, this toolkit provides linking functions for metadata (such as structural metadata) with resource files. Cataloger can finish the linking steps as constructing the resource structure, while creating each node, he can assign the linked resource files to the node. As the structure construction, page node resource linking can be done automatically without cataloger’s operation.

3.4 Importing Metadata

Sometimes we have some existing metadata records stored in other formats (such as those metadata records stored in TH-ADL[14]), or have some template input metadata files, to reduce the cataloger’s work, we provide metadata importing function in this toolkit. Using this function, we can import most administrative metadata items from some template files for a batch of digital resources of the same type, reducing the time of inputting data in hand. For those metadata records stored in other systems such as TH-ADL, metadata mapping should be made between the importing format and DC. Figure 2 shows the mapping table of the metadata between TH-ADL and DC.

3.5 Packaging and Creating METS Objects

After cataloging and resource linking, we can use this toolkit to package all types of metadata with linking information to create METS objects. The result objects are verified and ensure validation;
and ensure validation; documents are saved in standard METS format.

<table>
<thead>
<tr>
<th>TH-ADL</th>
<th>DC</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 作者信息</td>
<td>01 Creator</td>
</tr>
<tr>
<td>02 书名</td>
<td>02 Title</td>
</tr>
<tr>
<td>03 书名作者</td>
<td>03 Title.creator</td>
</tr>
<tr>
<td>04 出版单位</td>
<td>04 Publisher</td>
</tr>
<tr>
<td>05 年代信息</td>
<td>05 Date</td>
</tr>
</tbody>
</table>

**Figure 2.** Mapping table of metadata between TH-ADL and DC

### 4 Implementation

#### 4.1 Implementation Environment

We have developed the toolkit for TH-DLMS preservative digital objects creation. We use a PC server running Windows 2000 Advance Server to distribute the toolkit, main developing toolkit is JDK 1.4.2 or later for XML parsing and creation, we use JWSDP 1.2 (including JAXB 1.01) from Sun Microsystems.

#### 4.2 Using XMIL and Schema

We use XML as the final format of metadata storage, not to store metadata into RDBMS, because XML documents can carry data independent of system, and play an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere. For all types of metadata, we defined a schema for each, because schema is more flexible than DTD and suitable for defining the structure, content and semantics of XML documents.

#### 4.3 Using Java and JAXB

We use pure Java solution to develop this tool, so that this application can be ported to other platforms easily as needed. For XML processing, Java is more convenient and powerful than other languages. There are many kinds of Java-based XML processing APIs and toolkits, such as JDOM, DOM4J, JAXP JAXB, we choose JAXB as XML processing framework in our programs for its a high level interface for processing XML as Java Objects and easy to do validation.

**Figure 3.** An example of Chinese ancient book being cataloged

For simple images, metadata and digital images are submitted to the management module directly; while for books, metadata and digitalized images should be packaged into a whole _TEB_ (Tsinghua EBook standard) file extended from Open Ebook standard[15], then submitted to the management module. Figure 4 is a part example of a METS-encoded document for descriptive.

### 5 Results and Example

We have tested this cataloging toolkit for some kinds of digital resources, such as those simple image resources in TH-ADL[14], some digitalized Chinese ancient books, and Figure 2 is an example of a Chinese ancient book being cataloged. After cataloging, the metadata of digital resources are stored in standard METS XML documents, which are converted to Fedora-Extended METS documents later and submitted into the system.

### 6 Future works

At present, we have tested this toolkit for simple images and digitalized images for books; both of those resources can be
cataloged and saved into METS–encoded XML documents, which contain full metadata information for management and providing service. In next step, we plan to do test on other types of digital resources, such as audio and video resources, which have different administrative metadata and need different kinds of service, we should try to work out suitable administrative metadata standards devoted to the research and development of TH-DLMS, thanks them all for their efforts. Thanks to DENG Ke, who provides the solution for representing the electronic books; and thanks to WANG Yong, LI Huajing, who are responsible for studying Fedora and integrate it into TH-DLMS for digital objects management.

References
   http://www.loc.gov/standards/mets/
   http://www.fedora.info
   http://www.dublincore.org/documents/dces/
4. The Making of America.
   http://sunsite.berkeley.edu/MAOA/
   http://lcweb.loc.gov/rr/mopic/avprot/
6. Open Archives Initiative.
   http://www.openarchives.org/
   http://dlib.lib.tsinghua.edu.cn/tdlmstechmd.xsd
   http://dlib.lib.tsinghua.edu.cn/tdlmsrightsmd.xsd
   http://dlib.lib.tsinghua.edu.cn/tdlmsrcmd.xsd
    http://dlib.lib.tsinghua.edu.cn/tdlmsdigiprovmd.xsd
11. METS Implementation Registry
    http://sunsite.berkeley.edu/mets/registry/
12. Reference Model for an Open Archival Information System (OAIS).
15. Open eBook Forum.
    http://www.openebook.org/