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, dcszlz}@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 Source system[2].

Keywords Metadata Encoding and Transmission Standard (METS); Dublin Core[3]; XML.

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 full-featured 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.

(1) 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 official website (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 represents a node of the digital resource's structure.

(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.

2.2 Standard for 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 Archival Information System (OAIS) Reference Model[12].

2.3 Characters and Encoding

We choose Unicode for internal 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

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 in the result XML document.

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.

```
- <METS:structMap TYPE="Catalog">
  - <METS:div TYPE="book" ORDER="1" DMDID="DM1" LABEL="九章算术"
    ORDERLABEL="1">
    <METS:fptr FILEID="DS1" />
    <METS:fptr FILEID="DS476" />
   + <METS:div TYPE="chapter" ORDER="1" LABEL="序言" ORDERLABEL="1">
   + <METS:div TYPE="chapter" ORDER="2" LABEL="第一章" ORDERLABEL="25">
   + <METS:div TYPE="book" ORDER="3" LABEL="第二章" ORDERLABEL="75">
   + <METS:div TYPE="chapter" ORDER="4" LABEL="第三章" ORDERLABEL="125">
   + <METS:div TYPE="chapter" ORDER="5" LABEL="第四章" ORDERLABEL="175">
   + <METS:div TYPE="chapter" ORDER="6" LABEL="第五章" ORDERLABEL="225">
   + <METS:div TYPE="chapter" ORDER="7" LABEL="第六章" ORDERLABEL="275">
   + <METS:div TYPE="chapter" ORDER="8" LABEL="第七章" ORDERLABEL="325">
   + <METS:div TYPE="chapter" ORDER="9" LABEL="第八章" ORDERLABEL="375">
   + <METS:div TYPE="chapter" ORDER="10" LABEL="第九章" ORDERLABEL="425">
    </METS:div>
  </METS:structMap>
```

Figure 1. Structure representation of a book's content

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.

TH-ADL	DC
<ts:正题名>, <ts:副题名>	<dc:Title>
<ts:描述>	<dc:Description>
<ts:主要个人责任者>, <ts:主要团体责任者>	<dc:Creator>
<ts:次要个人责任者>, <ts:次要团体责任者>	<dc:Contributor>
<ts:论题主题>, <ts:题名主题>, <ts:中国法分类号>	<dc:Subject>
<ts:数字式资源出版者>, <ts:数字式资源出版地>	<dc:Publisher>
<ts:数字式资源制作日期>	<dc:Date>
<ts:资源类型>	<dc:Type>
<ts:资源格式>	<dc:Format>
<ts:URL>	<dc:Identifier>
<ts:资料出处 URL>, <ts:资料来源出版地> <ts:资料来源出版者>, <ts:资料来源出版日期>	<dc:Source>
<ts:语言>	<dc:Language>
<ts:关联资源>	<dc:Relation>
<ts:空间位置>, <ts:时间范围>	<dc:Coverage>
<ts:版权>, <ts:用户权限管理信息>	<dc:Rights>

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 XML 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 with different 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 it's a high level interface for processing XML as Java Objects, and easy to do validation of XML.

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 3 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 Fedora repository.

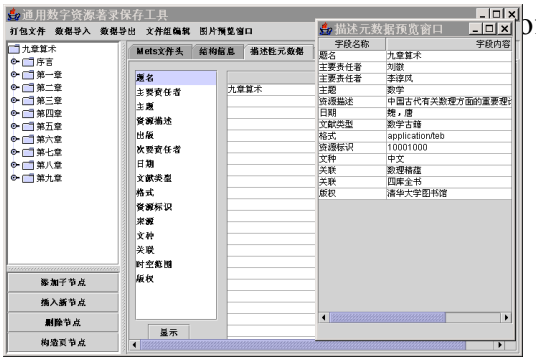

Figure 3. An example of Chinese ancient book being cataloged

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 metadata.

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 for them and add them to our proper services.

Acknowledgments. Many people have 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 ANG Yong, LI Huajing, who are responsible for studying Fedora and integrate it into TH-DLMS for digital objects management.

```
<METS:dmdSec ID="DM1">
  <METS:mdWrap MDTYPE="OTHER" MIMETYPE="text/xml"
               LABEL="清华大学通用数字资源描述元数据">
    <METS:xmlData>
      <oai_dc:dc xmlns:oai_dc=
      "http://www.openarchives.org/OAI/2.0/oai_dc/"
      xmlns:dc="http://purl.org/dc/elements/1.1/">
        <dc:title>九章算术卷一</dc:title>
        <dc:creator>刘徽</dc:creator>
        <dc:creator>李淳风</dc:creator>
        <dc:subject>数学</dc:subject>
        <dc:type>数学古籍</dc:type>
        <dc:description>中国古代有关数理方面的
        重要理论的著作</dc:description>
        <dc:format>application/teb</dc:format>
        <dc:identifier>10001000</dc:identifier>
        <dc:language>中文</dc:language>
        <dc:date>魏，唐</dc:date>
        <dc:relation>数理精蕴</dc:relation>
        <dc:relation>四库全书</dc:relation>
      </oai_dc:dc>
    </METS:xmlData>
  </METS:mdWrap>
</METS:dmdSec>
```

Figure 4. Part of a METS-encoded descriptive metadata

References

1. Metadata Encoding and Transmission Standard (METS). http://www.loc.gov/standards/mets/

2. The Fedora Project: An Open-Source Digital Repository Management System. http://www.fedora.info

3. Dublin Core Metadata Element Set, Version 1.1: Reference Description. http://www.dublincore.org/documents/dces/

4. The Making of America. http://sunsite.berkeley.edu/MOA2/

5. Digital Audio-Visual Preservation Prototyping Project. http://lcweb.loc.gov/rr/mopic/avprot/

6. Open Archives Initiative. http://www.openarchives.org/

7. XML Schema for Tsinghua University Digital Library Digital Image Technical MetadataDraft http://dlib.lib.tsinghua.edu.cn/dlms/th_dlmstechmd.xsd

8. XML Schema for Tsinghua University Digital Library Rights MetadataDraft http://dlib.lib.tsinghua.edu.cn/dlms/th_dlmsrightsmd.xsd

9. XML Schema for Tsinghua University Digital Library Source MetadataDraft http://dlib.lib.tsinghua.edu.cn/dlms/th_dlmssrcmd.xsd

10. XML Schema for Tsinghua University Digital Library Digital Provenance MetadataDraft. http://dlib.lib.tsinghua.edu.cn/dlms/th_dlmsdigiprovmd.xsd

11. METS Implementation Registry http://sunsite.berkeley.edu/mets/registry/

12. Reference Model for an Open Archival Information System (OAIS). http://ssdoo.gsfc.nasa.gov/nost/isoas/www classic/documents/pdf/CCSDS-650.0-B-13.pdf

13. Extensible Markup Language (XML) 1.0 Third Edition. http://www.w3.org/TR/2004/REC-xml-20040204/

14. Chunxiao Xing, Lizhu Zhou. A study and development of THADL digital library. Proceedings of the Sixth International Conference for Young Computer Scientist: in Computer Science and Technology in New Century, Hang Zhou, P.R., China, Oct 23-25 2001, International Academic Publishers, 2001, 605-609.

15. Open eBook Forum. http://www.openebook.org/