11-2002

A Metadata Framework Developed at the Tsinghua University Library to Aid in the Preservation of Digital Resources

Authors: Jinfang Niu

This article provides an overview of work completed at Tsinghua University Library in which a metadata framework was developed to aid in the preservation of digital resources. The metadata framework is used for the creation of metadata to describe resources, and includes an encoding standard used to store metadata and resource structures in information systems. The author points out that the Tsinghua University Library metadata framework provides a successful digital preservation solution that may be an appropriate solution for other organizations as well.

ARTICLES

D-Lib Magazine
November 2002

Volume 8 Number 11

ISSN 1082-9873

A Metadata Framework Developed at the Tsinghua University Library to Aid in the Preservation of Digital Resources

Jinfang Niu
Tsinghua University Library
Beijing 100084, P.R.China
<niujf@lib.tsinghua.edu.cn>

Abstract

This article provides an overview of work completed at Tsinghua University Library in which a metadata framework was developed to aid in the preservation of digital resources. The metadata framework is used for the creation of metadata to describe resources, and includes an encoding standard used to store metadata and resource structures in information systems. The author points out that the Tsinghua University Library metadata framework provides a successful digital preservation solution that may be an appropriate solution for other organizations as well.

Background

Digital preservation is an urgent problem for which solutions must soon be found. As a result, the focus of metadata research is increasingly moving from research on descriptive metadata to research on preservation metadata. In July 2001, the Goettingen State and University Library in Germany, Cornell University Library in the United States, Orsyell Library in France and Tsinghua Library in China agreed to collaborate on a project named EMANI. The EMANI project requires each of the four participants to develop a preservation system that will guarantee the long-term availability of their digital mathematic resources and that will enable the sharing of those resources with each of the other project participants. This article provides a description of system analysis completed at Tsinghua University Library as part of the EMANI project. We developed a metadata framework that can be used to adequately describe resources, as well as an encoding standard that can be used to store metadata and resource structures in information systems. The Tsinghua University Library metadata framework is the core and blueprint on which its preservation system for the Library's mathematics collection for EMANI will be built. One acknowledgement that needs to be made here is that our work at the Tsinghua University Library borrowed much from the Digital Audio-Visual Preservation Prototyping Project at the United States Library of Congress [1].

Metadata framework

The metadata framework we established at the Tsinghua University Library is module based. The framework includes a DMD module (descriptive metadata), a RightMD (rights metadata) module, a TechMD module (technical metadata), a SourMD (source metadata) module and a DigProMD module (digitization process metadata). The DMD module is an indispensable part of our preservation metadata scheme because resources in any collection will be inaccessible without descriptive metadata; RightMD is used for access control; and TechMD records the technical features of digital objects. (The preservation function of the metadata framework mainly lies in this module.)

The TechMD module has 5 sub modules: AllFileMD, TxtMD, ImagMD, AudMD, and VidMD. AllFileMD includes technical metadata common to all kinds of digital files. Each of the other 4 sub modules of TechMD relate to technical features specific to one of the following types of file: text files, image files, audio files or video files.

The SourMD module records the characteristics of source objects. The source object might be an analog object, such as a paper book, or it might be a digital file, such as a text, image, audio or video file, etc. When the source object is analog, the DigProMD module can be used to record the process of digitization. See Figure 1.

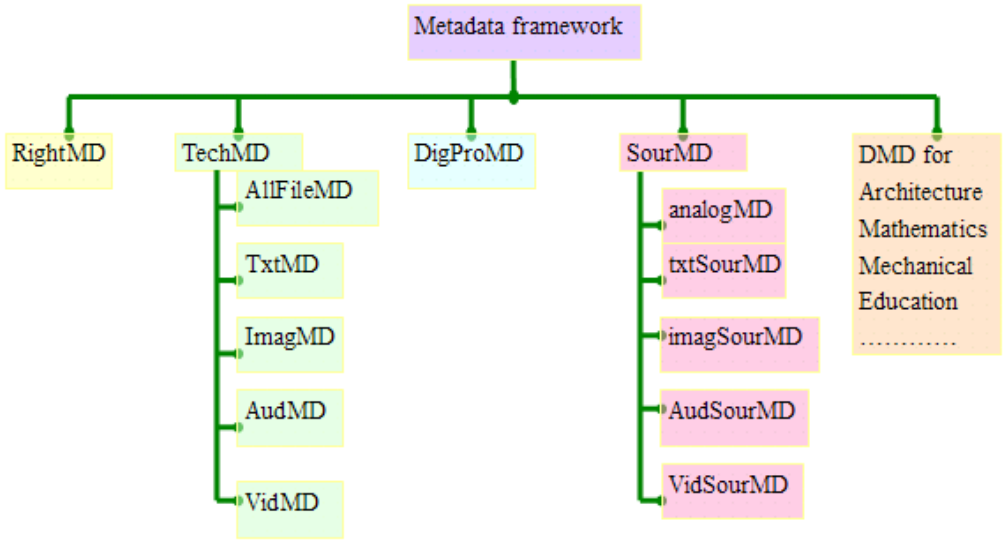

Figure 1. The Tsinghua University Library metadata framework.

Although we developed our metadata framework specifically for mathematic resources, the framework is applicable to all kinds of digital resources. When a resource changes, we change the descriptive metadata. Today, we are using the descriptive metadata for mathematic resources, but in the future we might use it for mechanics or education resources. To make the descriptive metadata for different resources consistent, we chose 12 of the most commonly used elements for the core metadata to describe Tsinghua University Library resources. All Tsinghua University Library metadata schemes will include these 12 elements (see Figure 2).

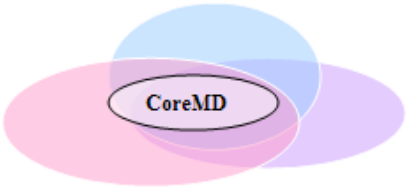

Figure 2. Core metadata as a part of a particular metadata scheme.

In our metadata framework, only the DMD module is indispensable; it provides the basic information necessary to certify the existence of resources. This requirement proves very useful when we want to upload lots of resources quickly but have inadequate time and labor to describe the resources in detail. The other four modules may be used for any kind of resources; however, they are optional. Thus, although the metadata framework seems large and complex, it is scalable. The metadata framework for a particular type of resource does not have to include all 15 unqualified DC elements. On the other hand, the framework can be made very complex when such complexity is needed. Catalogers decide whether the metadata framework for a particular type of resource will be simple or complex according to practical needs.

The metadata framework we developed can also describe the transformation history of resources (as illustrated in Figure 3). For born-digital resources, the metadata describes both the source digital file and current digital file. For a digitized resource, the metadata describes the analog object, the digitization process and the technical features of current digital files.

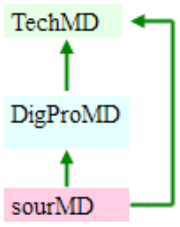

Figure 3. Relationship of modules used to describe the transformation history of resources.

Description mechanism

Structuring metadata properly is critical for providing accurate and easily understood resource descriptions. To fully represent the structure of resources, the metadata framework should be combined with a proper description mechanism. We have established a common structure for all kinds of resources, including books, websites, etc. We describe resources from the top layer to the lowest layer according to the structure for that resource. (See Figure 4.) Although only five layers are defined in the structure map shown in Figure 4, both the set and intermediate layers can embed lower layers of sets and intermediates, so actually, the structure is capable of accommodating many more layers than those five.

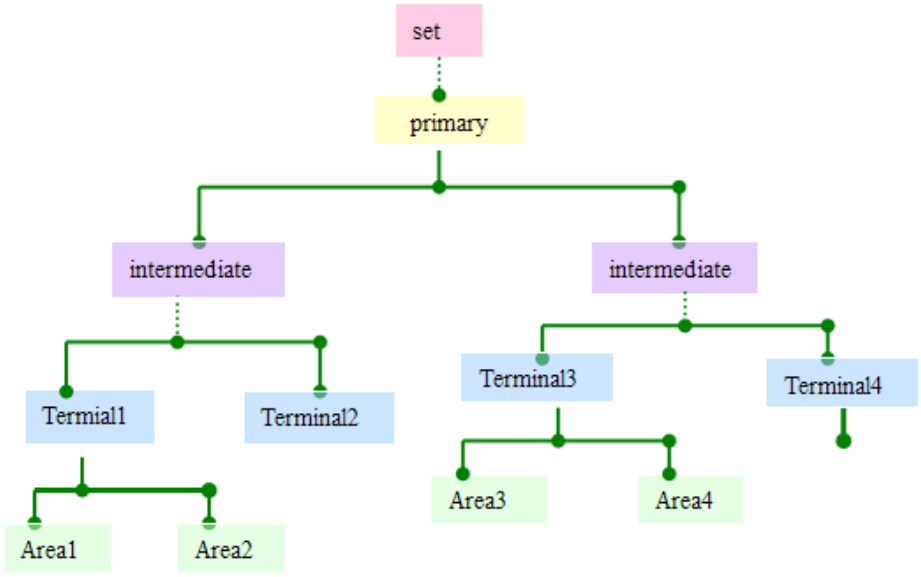

Figure 4. Illustration of the structural layers of a resource.

Following are definitions for the various structural layers shown in Figure 4 above:

- **Set:** a set is a congregation of many primary objects. Upper-layer sets can embed lower-layer sets. The number of layers embedded and the number of sets on one layer are unlimited. That is to say, there might be no set at all or many layers of sets. For example, *the resources collection of Hua Luogeng,* a famous mathematician, might include a *letters collection* and a *manuscripts collection.* According to the particular features of resources, as well as practical requirements, catalogers decide what number of set layers may be needed.

- **Primary:** A primary object is one that has independent and complete content, such as a book, a website or a film.

- **Intermediate:** An intermediate layer is a part of a primary object, but it also includes a number of terminal objects. Like the set layer, sometimes the intermediate layer doesn't exist. Sometimes there are several lower-layer intermediate objects embedded in an upper-layer intermediate object. The number of intermediate layers is determined by catalogers according to resource features and practical needs. For example, when a book consists of 3 volumes and each volume is stored as only 1 file, then the number of intermediate layer objects is zero. The structure of the book is: a *primary* object consisting of 3 *terminal* objects. However, when a book has a structure like is shown in Figure 5, that primary object might have 3 layers of intermediate objects consisting of the volume, the chapter, and the page.

- **Terminal:** A digital file is defined as a terminal object.

- **Area:** An area is a sub unit of a terminal object. Under some circumstances, a file may include many relatively complete or important parts, for example, a CD named *songs of Wangfei* may have ten songs in one digital file. In that case, each sub unit—representing a song—can be regarded as one area.

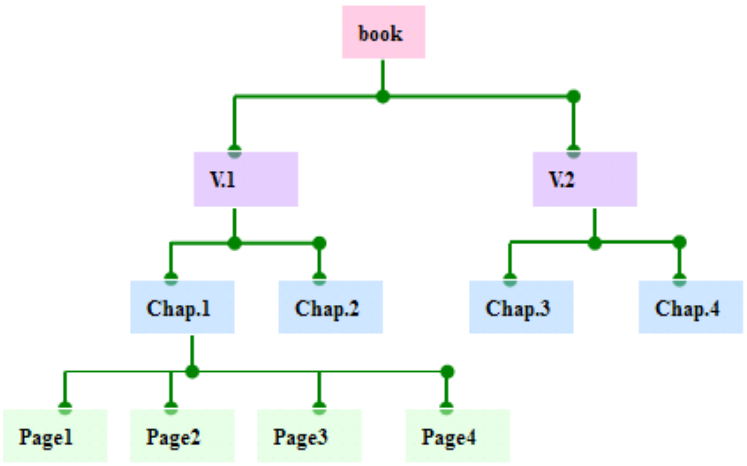

Figure 5. An example of the various layers of a particular book.

When we describe resources using the metadata framework, different metadata modules are applied to different layers in the structure. Objects in a lower layer inherit the metadata of upper layers. Descriptive metadata are used only for primary objects. Technique metadata are used only for terminal objects. According to practical needs, catalogers decide which metadata modules should be used within the different layers of the structure for a resource. For example, if the whole set of resources have identical rights information, the RightMD is used within the set layer only. When Chapter 1 and Chapter 2 of a book have different rights information, the RightMD should be used at the layer describing the two chapters respectively.

In our metadata framework we provide structural description interfaces according to the common structure. After the person creating the metadata completes the upper layer description, he or she presses a button, and then enters the next interface to describe the lower layer objects. Figure 6 is an illustration of the description interface of the primary layer.

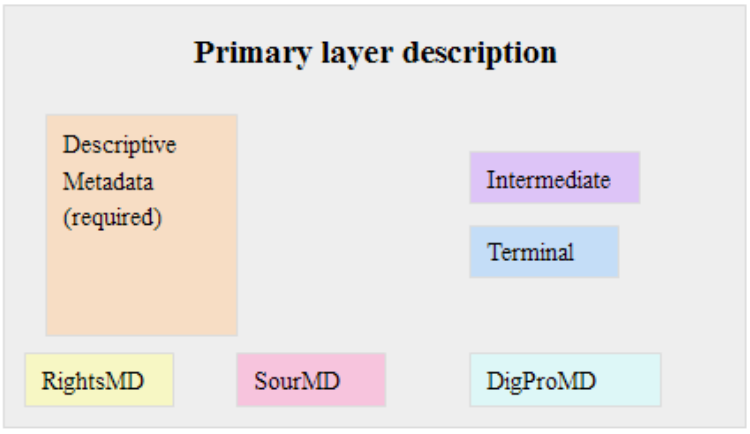

Figure 6. Illustration of the primary layer description.

Metadata Encoding

The metadata framework we developed, combined with the description mechanism, fully and adequately describes resources. The next step is to store metadata and structure in an information system. We chose METS as our encoding standard. METS (Metadata Encoding and Transmission Standard) [2] was developed by the Digital Library Federation [3] and is maintained in the Network Development and MARC Standards Office [4] of the United States Library of Congress. METS provides a standard way to represent resource structure and a standard method to encode metadata (using XML). The METS structure works well with our metadata framework and resource structure. METS has 4 main parts:

- dmdSec (descriptive section),
- amdSec (administrative section),
- fileGrp (list and group all the files contained in one description object), and
- structMap (represent the structure of the description object).

Descriptive metadata in our metadata framework are encoded in dmdSec. RightMD, techMD, sourMD and digiproMD are encoded amdSec. All the terminal objects (files) are listed in the fileGrp section. The hierarchy structure of the description object is encoded in the structMap section. Following is a simplified example.

```
<METS:mets>
<METS:dmdSec ID="DMD1">
    <METS:mdWrap >
        <METS:xmlData>
            <dmd:dmd>
                <dmd:title>九章算术</dmd: title>
             </ dmd:dmd >
         </METS:xmlData>
      </METS:mdWrap>
</METS:dmdSec>
<METS:dmdSec ID="DMD2">
    <METS:mdWrap >
        <METS:xmlData>
            <dmd:dmd>
                <dmd:title>孙子算经</dmd: title>
             </ dmd:dmd >
         </METS:xmlData>
     </METS:mdWrap>
 </METS:dmdSec>
<METS:amdSec>
     <METS:TechMD ID="ADM1">
         <METS:mdWrap >
             <METS:xmlData>
                 <adm:AllfileMD USE="optional">
                     <adm:fileID>file1</adm: fileID >
                     <adm:location> tsinghualib-sxgj_ijss_1</adm: location>
                  </adm: AllfileMD>
              </METS:xmlData>
          </METS:mdWrap>
```

```
        </METS: TechMD>
<METS:TechMD ID="ADM2">
      <METS:mdWrap >
         <METS:xmlData>
              <adm:AllfileMD USE="optional">
                    <adm:fileID>file2</adm: fileID >
                    <adm:location> tsinghualib-sxgj_ijss_2</adm: location>
                </adm: AllfileMD>
            </METS:xmlData>
      </METS:mdWrap>
   </METS: TechMD>
<METS:TechMD ID="ADM3">
      <METS:mdWrap >
         <METS:xmlData>
              <adm:AllfileMD USE="optional">
                    <adm:fileID>file3</adm: fileID >
                    <adm:location> tsinghualib-sxgj_ijss_3</adm: location>
                </adm: AllfileMD>
            </METS:xmlData>
      </METS:mdWrap>
   </METS: TechMD>
 </METS:amdSec>
<METS:fileSec>
    <METS:fileGrp VERSDATE="2002-07-03">
         <METS:fileGrp>
              <METS:file ID="File1" SEQ="1" DMID="ADM1" >
                  <METS:Flocat xlink:href="tsinghualib-sxgj_ijss_1" />
              </METS:file>
              <METS:file ID="File2" SEQ="2" ADMID="ADM2" >
                  <METS:FLocat xlink:href="tsinghualib-sxgj_ijss_2" />
              </METS:file>
              <METS:file ID="File3" SEQ="3" ADMID="ADM3" >
                  <METS:FLocat xlink:href="tsinghualib-sxgj_szsj_3" />
              </METS:file>
         </METS:fileGrp>
    </METS:fileGrp>
</METS:fileSec>
<METS:structMap>
<METS:div ORDER="1" LABEL="清华图书馆数学古籍" >
      <METS:div ORDER="1" LABEL="九章算术" DMDID="DMD1">
          <METS:fptr FILEID="File1" />
          <METS:fptr FILEID="File2" />
        </METS:div>
        <METS:div ORDER="2" LABEL="孙子算经" DMDID="DMD2">
          <METS:fptr FILEID="File3" />
        </METS:div>
    </METS:div>
```

```
</METS:structMap>
</METS:mets>
```

Comparing with other standards

Dublin Core (DC) [5] is mainly a descriptive metadata set. It aims to facilitate resource discovery of digital resources. It doesn't record the transformation history of resources from analog to digital nor, in the author's opinion, does it include enough technical metadata to guarantee long-term preservation. DC does have one element that can describe structure. That element is the DC *relation* element with its two sub-elements: *has part* and *is part of*. However, the *relationship* they can describe is too simple. Using the *relation* element allows you only to see one layer above and one layer below the current layer. You can't view the entire hierarchy structure at once. If you want to view the entire structure, you have to look for it by following the relation link across many DC records. For the structure illustrated in Figure 7, for example, you have to synthesize the information recorded in the following four records:

Record 1: Chapter1 *is part of* volume1

Record 2: volume1 *has part* chapter1 and chapter2 volume1 *is a part of* book1

Record 3: book1 *has part* volume1 and volume2

Record 4: volume2 *has part* chapter3 and chapter4 volume2 *is a part of* book1

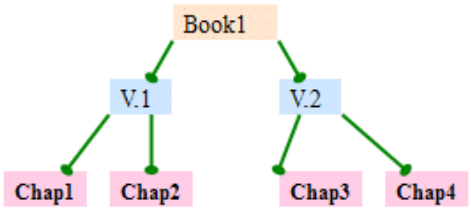

Figure 7. Illustration of the structural layers of a book.

For a book such as an academic textbook, the structure might be much more complex, involving the synthesizing of many more records. It's vry inconvenient for customers to get an idea of the whole structure. Although the core metadata in our description module is based on DC, in fact DC could be only one of the metadata modules we might decide to include in our metadata framework.

The CEDARS [6] and NLC (National Library of China) [7] preservation metadata schemas incorporate DC as a part of their metadata frameworks and serve the functions of long-term preservation and resource discovery. However, like DC, their metadata frameworks don't provide mechanisms to represent resource structure.

Previously, metadata researchers focused their attention on metadata element sets. Many metadata elements sets are widely discussed and some are even set as national standards, such as DC. However, although structure is a very important characteristic of resources, mechanisms to describe and represent resource structure have been ignored until now.

Since there are no widely known standards to follow for representing resource structure, metadata designers and system engineers have had to develop their own methods. As different people develop different methods to suit their particular situations, difficulties for information exchange and system interoperation are the inevitable result. OEB (Open eBook™ Publication Structure) [8] is the first standard—as far as we know—that combined DC metadata with a structure representation mechanism. It's a very good standard for electronic books. However, according to our analysis, the scheme we developed at the Tsinghua University Library surpassed OEB in at least two aspects:

1. The structure we defined is applicable to all kinds of resources, such as books, websites, collections and any other resources that have a hierarchical structure. OEB only deals with the structure of books.

2. OEB doesn't focus on digital preservation, so it only includes descriptive metadata.

Conclusion

Members of the EMANI workgroup at Tsinghua University Library agree that the metadata framework we developed is useful and sufficient for preserving digital resources and representing resource structure. We will present our metadata framework to the other three parties of the EMANI project in the near future. If the other EMANI project participants approve our metadata framework, we will begin system development based on that framework. The resulting preservation system will accommodate the metadata framework, adopt the METS encoding standard and provide the structural description mechanism.

The Tsinghua University Library metadata framework enables resource structure to be fully described and represented, and the digitization process from analog to digital can be clearly described and organized. The universality and flexibility of our metadata framework is also evident. For that reason, the framework we developed may be appropriate for other organizations as well.

Acknowledgements

I gratefully acknowledge my director, Airong Jiang, who encouraged me to publicize our work on digital preservation. I also thank Xiaohui Zheng and Ting Zeng, who participated in the discussion about the metadata framework, the XML schema and the description mechanism.

References

[1] Digital Audio-Visual Preservation Prototyping Project at the United States Library of Congress. Available at <http://lcweb.loc.gov/rr/mopic/avprot/avprhome.html>.

[2] Metadata Encoding and Transmission Standard (METS). Available at <http://www.loc.gov/standards/mets/METSOverview.html>.

[3] The Digital Library Federation home page <http://www.diglib.org>.

[4] The Network Development and MARC Standards Office, U.S. Library of Congress, <http://lcweb.loc.gov/marc/ndmso.html>.

[5] Dublin Core home page <http://dublincore.org/index.shtml>.

[6] Cedars Project home page <http://www.leeds.ac.uk/cedars/>.

[7] National Library of China home page <http://www.nlc.gov.cn>.

[8] Open eBook™ Publication Structure
<http://www.openebook.org/oebps/oebps1.2/index.htm>.

DOI: 10.1045/november2002-niu