The Digitization Project of the Vatican Library within the Complex Relationships between Sets of Metadata

Paola Manoni

Biblioteca Apostolica Vaticana
manoni@vatlib.it

Abstract. The talk is focused on the metadata schemas involved in the digitization project of the Vatican Library: for long-term preservation strategies as applied to digital deposit collections, as well as for web-publication of images in the context of the digital library. The relationship management in the implementation process of sets of metadata (in their structural representation and semantic meaning of data elements) will be discussed with particular attention to the management implications and the resulting operational capabilities.

Keywords: Metadata, Digital Libraries, MARC21, TEI-P5, METS, PREMIS, FITS format.

1 Introduction

Metadata is the core of any information retrieval system and so its implications for any digital library are profound: the choice of a metadata scheme underpins any such library's ability to deliver objects in a meaningful way, and greatly affects its long-term ability to maintain and preserve its digital assets.

The overall goal of this presentation is to provide information about metadata infrastructures that affords interoperability among heterogeneous, autonomous digital library services implemented for the digitization project of the Vatican Library. These services include both search services and remotely usable information processing facilities.

Metadata required for a diverse set are surveyed and classified. Metadata architecture fits into our established infrastructure and promotes interoperability among existing and de-facto metadata standards. Several pieces of this architecture are implemented; others are under construction. The architecture metadata information offers facilities for search services, and local metadata repositories. In presenting and discussing the pieces of the architecture, we show how they address our motivating requirements. Together, these components provide, exchange, and describe metadata for information objects and metadata for information services.

2 Metadata Related to Objects

The digitization is performed for various projects operating in the library that relate to collections of manuscripts and incunabula.

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 1–6, 2014.

Descriptive metadata are primarily used for resource discovery and format currently used include TEI-P5 format for manuscripts and MARC21 for incunabula.

The following summary describes metadata protocols involved in the interoperability between systems being implemented as **research tools** of the Vatican Library and the workflow planned for the forthcoming digital project.

3 Brief Description of the Existing Research Tools

3.1 Online Catalogues

The Vatican Library provides online catalogues to help researchers access and make better use of the collections.

There are separate online catalogues for each collection (Manuscripts, Archives, Printed books, Incunabula, Graphic prints and Drawings, Coins and Medals) and a General Catalogue able to perform interoperability between MARC21, EAD and TEI-P5.

- **Manuscript catalogue**: It includes complete or partial data taken from inventories, bibliographies, printed catalogues, card indexes. The encoding of the descriptive elements conforms to the **TEI specifications** and uses XML syntax;
- **Archival holdings catalogue**: It includes complete or partial data taken from inventories. The encoding of the descriptive elements conforms to the **EAD specifications** and uses XML syntax;
- **General Printed books catalogue**: It includes the description of the entire collection of printed volumes (monographs and periodicals) from the XVIth century to the new acquisitions. Cataloguing is carried out in MARC 21 format;
- **Incunabula catalogue**: It includes bibliographic records related to the VISTC (*Vatican Incunabula Short Title Catalogue)* and the *BAVIC*[1] of the entire collection where links to persistent URIs related to the digitized volumes are provided. Cataloguing is carried out in MARC 21 format;
- **Graphic prints and Drawings catalogue**: It includes the descriptions of the prints, maps, drawings, photographs and plates which are kept in the various collections of the Library. Cataloguing is carried out in MARC 21 format;
- **Coins and Medals catalogue**: It includes descriptions of the coins and medals kept in the Library. A running project aims to make digital scans of the graphic prints, coins and medals, insert hypertextual links to them into the related bibliographical descriptions, and enter the data into each web-based catalogue of the Library Cataloguing is carried out in MARC 21 format.

3.2 Systems in Use

The online catalogues use two main systems for the management of different metadata: TEI-P5 and EAD in XML syntax for manuscripts and archival units (in two

[1] The project BAVIC (Bibliothecae Apostolicae Vaticanae Incunabulorum Catalogus) is the analytical cataloging of 8,600 incunabula.

separate collections of data but in the same application named InForMA, entirely developed at the Vatican Library using open-source Java/XML technology for data archive, authority indexes and search engine); and MARC21 for the other collections (with different specifications for each type of document).

- **InForMA** — The implementation provides: full native XML database support, storing XML content as is and providing true XML retrieval capabilities based on the XPath and XQuery standards; handling of structured and unstructured multimedia rich content (it can store and retrieve multiple different file types); integration with Microsoft Office and other WebDAV-enabled product suites. These features are supported by the embedded **Tamino** technology provided by the German Software AG company.

  The implementation of the TEI schema for manuscripts also provides a kind of information that includes data element as atomic units applied for internal usage in which the persistent URI related to the web-presentation for each digitized manuscript is given.

- **V-Smart / Iguana** – It allows scholars to query either the integrated general catalogue, where they can have a quick and thorough search result related to any bibliographic resource, or each catalogue where bibliographic descriptions are stored in their native format. The technological infrastructure is based on OAI-PMH protocol and script elements for exporting data (from InForMA) and importing XML (to V-Smart). From each bibliographic record, V-Link (openURL resolver) is able to search and access a variety of information resources and retrieve truly relevant search results.

  The system is able to support the persistent URIs related to the web-presentation of the digitized incunabula.

4 Structural Metadata

Information necessary to record the internal structure of an item so that it can be rendered to the user in a sensible form. This type of metadata is necessary as an item may often be comprised of multiple of the images of individual pages that make up a digitized book.

For the digital project the Vatican Library has adopted METS standard automatically created by the use of the DWORK[2] implementation, provided by the UniversityLibrary of Heidelberg to support the process of digitization and the web presentation of the digital objects.

[2] The University Library of Heidelberg in-house development. It supports the process flow of digitization and the web presentation of the digitized works.

The software as a web application thereby supports all single steps of the workflow from the creation of metadata, scan processing, creation of the web presentation to the storage of scans and metadata.

5 Administrative Metadata

They provide information about the management of the digital collection and facilitates long-term management and processing of it. They include:

- quality control, rights management, access control and use requirements;
- technical data on creation (such as scanner type and model, resolution, bit depth, color space, file format, compression, light source, owner, copyright date, copying and distribution limitations, license information, preservation activities (refreshing cycles, migration, etc.);
- preservation, action information.

While the first point in list pertains to the METS file generated for each digitized unit/volume, preservation and action information are managed within the PREMIS framework.

The use of the PREMIS is closely related to the concept of long-term preservation that the Library has adopted.

The photo workflow includes the acquisition of images in tiff/ raw formats. These files are used for the generation of the web-presentation procedure in the DWORK (METS and set of JPEG files). Then images are converted into FITS format and stored into the certified WORM data storage device for long-term preservation[3].

In the context of a discussion about metadata and how they interact, the consideration on the choice of storage formats has no place but, in the use of PREMIS in the Vatican Library, it is important to highlight the link between the information concerning the technical metadata format of the converted images and the use of the extension container 'object Characteristics Extension' that gives a place to record technical metadata defined by the FITS dictionary.

In fact a FITS file is made of 2880-byte records called 'FIT*S* blocks' divided between a header and a data area.

Each FITS file consists of one or more headers containing ASCII card images (80 character fixed-length strings) that carry keyword/value pairs, interleaved between data blocks. The keyword/value pairs provide information such as size, origin, binary data format, free-form comments, history of the data.

In the occurrence of the long-term archiving of the digital image, an automatic procedure extracts the embedded data in the header and defines the above mentioned

[3] Manuscripts and antique books of the Vatican Library's collections are being digitized to preserve them for future generations adopting a file format developed in the context of space missions and storing satellite images of the sky during the end of 70s of the last century: the Flexible Image Transport System (FITS), an open format, fully documented, without royalties or copyright, based on a series of specification publicly available and managed by a non-profit scientific authority.

The conversion TIFF/FITS takes into account all the characteristics of technical metadata (for example in TIFF format) in order to identify matches in the keywords of the header portion of the FITS file.

extension of the PREMIS. The structure of PREMIS is then completed with the main features related to the digital object and semantic units are defined to record the *environment* of each object.

6 Interoperability

Describing a resource with metadata allows it to be understood by both humans and machines in ways that promote interoperability. Interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality. Using defined metadata schemes, shared transfer protocols, and crosswalks between schemes, resources across the network can be searched more seamlessly.

The interoperability and exchange of metadata is further facilitated by metadata crosswalks. A crosswalk is a mapping of the elements, semantics, and syntax from one metadata scheme to those of another.

The workflow management of metadata, involved in each phase, from the acquisition of images to the archiving of the digital object, includes several steps and specific softwares:

1. Filename of TIFF / RAW file: An application able to assign a highly structured filename has been implemented[4].
2. The file name is automatically interpreted by the above mentioned DWORK so that the logical and physical sequences for each file group are added in the METS file related to each unit/volume.
3. Structural metadata that contain a table of contents with links to key structural elements such as title pages, table of contents, chapters, parts, sections and subsections (depending on the item) are added and automatically converted in the METS file.
4. Essential descriptive metadata are given in the MODS section of the METS file. Crosswalks between MODS and TEI-P5 / MARC21 has been established.
5. METS files related to manuscripts are exported from the DWORK to be processed by a specific console application in InForMA where the URI of the web presentation in the element 'FLocat LOCTYPE'is treated as an *ad hoc* element TEI-P5 document through specific Xquery functions.
6. METS related to incunabula are exported from DWORK in order to extract the 'FLocat LOCTYPE' for the creation of the link in the above mentioned OPACs, for each MARC21 bibliographic record. The information about the description of the digital object is performed in a qualified DC record in the OPAC.
7. TIFF format is converted in FITS format and a crosswalk between embedded technical metadata of TIFF and keywords in FITS header has been implemented.
8. PREMIS entities described in XML files are added at the time of archiving digital objects in the WORM storage device.

[4] Programming was carried out by specialists of the company Metis Systems s.r.l.

9. End-users can query the Web OPACs and get information about digitized manuscripts and incunables or browse the list of shelfmarks for each digital collections available in the website.
10. Reproductions of digital object may be requested. Queries are performed in the PREMIS database and a conversion from FITS file to an exchange format is available for private study or professional use.

References

1. Manoni, P.: Metadata framework and application profiles in the global structure of catalogs and digitization projects of the Vatican Library. Global Interoperability and Linked Data in Libraries: Special Issue. JLIS 4(1) (2013), http://leo.cilea.it/index.php/jlis/issue/view/536
2. Manoni, P.: L'interazione tra banche dati e analisi dei modelli descrittivi nella Biblioteca Apostolica Vaticana. In: Il Libro Antico Tra Catalogo Storico e Catalogazione Elettronica. Convegno internazionale, vol. 127. Accademia nazionale dei Lincei, Roma (2012)
3. Manoni, P.: The Vatican Library Web-based application for managing manuscript metadata stored in native XML databases. In: Current Research in Information Sciences and Technology, Proceedings of the Ist International Conference on Multidisciplinary Information Sciences and Technologies, Badajoz, Merida, vol. 2, pp. 570–575 (2006)