# Towards a sound management of digital culture. Metadata schemes and application profiles for digital repositories

**Pierluigi Feliciati**
University of Macerata – Department of cultural heritage (IT)
pierluigi.feliciati@unimc.it

**Abstract** – *If we give a look to the present panorama of digital research – particularly if applied to cultural heritage collections - we have to admit that there is still more attention to the quality of data than to the necessary series of management activities, extended to the long-time preservation of what was created (often spending a lot of money). Cultural heritage digital collections have to be considered in any sense digital libraries, even if they are not conceived to be delivered to web users. To manage the life-cycle management of digital libraries administrative and preservation metadata are widely considered to represent an essential tool. On this attention to metadata management should be based the distinction between projects oriented to the mere production of data and those oriented to service delivery (for closed or open communities). A short survey on metadata landscape is proposed, putting in evidence their advantages and issues, with respect to efficiency to inter-operate, manage and preserve what we can define the artifacts of our digital culture. The paper closes with an in-depth presentation of Italian experience, in particular with the MAG metadata application profile and its applications.*

**Keywords –** *Metadata, digital library, cultural heritage, digitization management, digital preservation, MAG*

## 1. INTRODUCTION

Professional networks concerning digital libraries, last years, defined some reference and application models based on the axiom that digital projects need management, even more than analogical ones.

In this framework, a sort of new *de facto* distinction emerged between digital collections conceived inside the heterogeneous and wide library environment (both reproducing analogical documents and aggregating digital born ones) and those not devoted to build new services, even inside the cultural heritage world. For example, often the archaeological or art pieces documentation or 3d modeling applications don't give enough importance to the definition of a policy for the long-term management of resulting digital resources.

The basic foundation of this paper is that cultural heritage digital collections have to be considered digital libraries, even if they are not conceived to be accessible to generic remote users. Moreover, digital resources resulting from any kind of project related with cultural heritage could be considered heritage themselves, coming after the commitment of substantial scientific, cultural and financial resources so worthy to be preserved.

In other words, those resources, built to offer a service for final users or for professional communities should be available also for future use(r)s. Thus, their management and long-term preservation emerge to be essential tasks, to be considered inside a specific policy, defined in the starting phases of the process requested by any project.

This process' life-cycle management covers both the organizational and the technical points of view. In particular for the second challenge, administrative and preservation metadata are widely considered to represent an essential tool to guarantee a sound management.

## 2. A MANAGEMENT POLICY FOR DIGITAL CULTURAL HERITAGE PROJECTS

In order to move this paper from some basics, we have to consider a possible definition for a digital library: first of all, this entity is not just a collection of digital documents but, for example “a (potentially virtual) organization that comprehensively collects, manages, and preserves for the long term rich digital content and offers to its user communities specialized functionality on that content, of measurable quality, and according to prescribed policies” [1].

Some others were even more radical in defining a digital library. The first principle of the *Digital Libraries Manifesto* published by the Study group on digital libraries of the Italian Association of Librarians states that “Digital libraries are conversations”, not “a single system or grand systematic narrative” [2]. The interactions (between resources and users and between users) was affirmed to be a crucial point. Anyway, in this paper we'll not treat such issue.

Inside the digital libraries' implementation and management activities, an essential part is

represented by its long-time preservation, i.e. "all activities concerning the maintenance and care for/curation of digital or electronic objects, in relation to both storage and access" [3] or "the act of maintaining information, in a correct and Independently understandable form, over the Long Term" [4].

What does mean it long-term preserving the digital resources resulting from a digital heritage project?

The starting step is to take always into account that a digital resource is inseparably composed of content (a sequence of bits) and a set of information (metadata), in order to make that sequence significant, identifiable and accessible for the use, storage, preservation, dissemination and for all other management operations. This metadata are more and more recognized as crucially important, regarded as a forming part of the very definition of a digital item not only in present but in its changing dynamics in times and spaces.

The function of normalizing the digital content metadata is also provided to support the automation of the digitization process, and helps to create a industrial market for quality products and services in this area.

Thus, an essential part of digitization projects – particularly those focused on cultural heritage [5] - consists in the accurate definition of one or more metadata sets associated with objects that will be part of a digital collection.

On this attention to metadata management should be based the distinction between projects oriented to the mere production of data and those oriented to service delivery. The first kind of projects create digital objects mostly with the goal of optimizing the cultural data analysis process, reducing the use of original analog documents or to obtain their copies, and they are often materialized in the production of large amounts of media not that easy to handle and manage, such as CD -ROM, DVD, DAT etc..

The latter class of projects take account of - and assume the responsibility to - certify the integrity of the information content and its storage conditions during the entire life-cycle, in order to ensure accessibility in the long term to a designed community of users. Whether they are mainly oriented towards the permanent preservation and accessibility of information, a consistent use of metadata favors the projects that ensure a 'total quality' of digital information and gains a more positive support in raising the necessary funding to support those long-term operations.

I'm convinced that this distinction has to be crossed over, because in a wider and up-to-date view every aggregation of digital data for every scope could be considered a service (a digital library?), made of resources and users and by their interactions. Users, in this sense, are not just those persons (“final” and remote) who access digital resources by the Web, but they are also data administrators, content professionals and even software/hardware agents.

## 3. THE MANAGEMENT METADATA LANDSCAPE

What are metadata in a DL management framework? They play the role of Pollicino, Kleinduimpje, Hop-o'-My-Thumb' pebbles: when abandoned by his parents, he finds a variety of means to save his life and the lives of his brothers... he drops the pebbles behind, discovering along the way that they‘re better than breadcrumbs to find back the path!

In other words, metadata seem to be the best solution to ensure the management of digital information over time, remembering the risk to lose digital information after a decade or even less: preservation of digital information is widely considered to require more constant attention than preservation of other media, such as built, written or painted heritage.

The creation and organization of metadata – even before digital era - has always been central in the activities of memory institutions (archives, libraries, museums, audiovisual centers), providing description of information resources (i.e. catalogs) to ensure their identification and retrieval, to fix documents relationships within and among objects or to manage resources over space and time.

To propose a classification of metadata, several taxonomies have been proposed. An interesting and easy-to-use typological document was published some years ago by the University of Melbourne [6] that proposed some possible oppositions: metadata can be *general* or *specialist*, *minimalist* or *rich*, *hierarchical* or *linear*, *machine generated* or *human authored*, *structured* or *unstructured*, *embedded* or *detached*, or they can be represented by *surface information* or even by *keywords, Google use of words, tags, user assigned infos*.

One of the most popular metadata classifications, with the advantage of simplicity and clarity, was the Wendler taxonomy [7], with metadata divided into three functional categories:

- *Descriptive*: to identify and recover digital objects; consisting of standardized descriptions of source documents (or documents digital natives) usually reside in the databases of information retrieval systems outside of the archives of digital objects, and are connected to them by links;
- *Administrative and management*: for the various management operations on digital objects within the archive; This may include technical informations about the digital objects creation, their storage format(s), copyright and licensing informations, and information necessary for the long-term preservation of the digital

objects.

- *Structural*: to describe the internal structure of documents (e.g., introduction, chapters, index of a book) and/or manage the relationships between various components of several related objects.

With a parallel approach, the NISO guide on metadata distinguished them in three classes: descriptive, structural and administrative. With structural metadata they mean a description of how the components of the object are organized, and administrative metadata are sub-divided into rights management and preservation [8].

The category of *preservation metadata* is related to those informations applicable to preservation actions: technical data on the format, structure and use of the digital content, the history of actions performed on the resource, the authenticity information such as technical features or custody history, and the responsibilities and rights informations.

Another important classification [9] focused on the role of metadata for data base implementation, distinguishing between *structural/control* metadata and *guide* metadata. The first class is used to describe the structure of computer systems such as tables, columns and indexes, the second conceived to help humans find specific items and is usually expressed as a set of keywords in a natural language.

Anyway, it's usual that the category of structural metadata is included in that of administrative ones, while the distinction between technical and administrative metadata is light: both categories help us to leave the right informations along the paths, and to build on them a long-term management policy.

The large variety of schemes available and the frequent overlapping of functions between metadata standards generates an intense crosswalk or mapping activities. The main issue is that there are not 100% perfectly equivalent metadata schema, semantically, in richness or granularity. Thus, during crosswalks it's usual to build many-to-one mapping rules, to force element's meanings or even to lose data. Thus, one of the main policy issues is to ensure the full scalability and interoperability to metadata schemes. Most of schemes and/or application profiles, anyway, are XML-based, making easier their possible interoperability.

A huge part of metadata standards activities was sustained by the Library of Congress [10], starting from the MARC bibliographic family of standards to the metadata schemes: *descriptive*, like MODS or MIX, *administrative* and *structural* like METS, for the *preservation* like PREMIS. In particular this last project, sponsored by OCLC and RLG from 2003-

2005 and then maintained from LOC, is focused on a *PREservation Metadata: Implementation Strategies*, by the definition of a general model and of a data dictionary, containing “a core set of semantic units that repositories should know in order to perform their preservation functions” [11].

An important role in metadata definition is played also by other organizations and working groups, for example by the Moving Picture Experts Group (MPEG) formed by the ISO to set standards for audio and video compression and transmission. They defined in particular the MPEG21XML-based standard, an open framework for multimedia applications, whose second part provides a *Digital Item Declaration Language* (DIDL), an interoperable schema for declaring the structure and makeup of what they call *Digital Items* [12].

Some among those management metadata application profiles – especially those released and maintained by important institutions like LOC with the main goal to recover and manage many resources coming from different sources - are conceived as powerful “packaging schemes” not providing direct solutions to specific scenarios. This involves a necessary and weighty activity of crosswalk for each exchange of data ad metadata.

Some other projects face this issue by defining “closed” Application Profile or Schema, public and documented but not as much open. They package standard *XML namespaces* and schema with scenario-formed elements, in order to answer to single, defined application scenarios.

## 4. THE ITALIAN EXPERIENCE WITH ADMINISTRATIVE NMETADATA

The Italian huge project of *Biblioteca Digitale Italiana* (Italian Digital Library) started by this last requirement, to ensure the production and aggregation - at national level and by many organizations - of many digital collections technically homogeneous.

The MAG (*Metadati amministrativi e gestionali*) application profile [14] was defined in this framework: totally compliant to international standards, allows the use of metadata maintained and defined in other schema (Dublin Core and NISO) in association with specific metadata defined for its particular scenario (just where we couldn't find a strengthened correspondence with existing schemes). It was conceived with the main goal of promoting among Italian cultural organizations the aggregation of a common set of technical and management metadata to guarantee the good submission and transfer of metadata and cultural digital objects (text, images, audio, video) in local or distributed digital libraries (SIP and DIP phases of OAIS model). In particular, it was conceived inside a national digitization project, not to manage digital-born documents.

The MAG metadata profile, expressed in XML, was conceived as an open standard, documented, freely available and completely independent from

specific hardware and software platforms.

To guarantee the support to MAG adoption activities, the AP is maintained since 2001 by a Committee supported by the ICCU – the Italian Central Institute for the Union Catalogue of Italian Libraries and Bibliographic Information, composed by experts from different fields: archives, libraries, human informatics, audiovisual, art objects 15]. The documentation of the 2.0.2 version presently includes a reference document in Italian and English [16], an Italian Handbook printed or digital and some examples of implementation.

MAG provides a formal specification for the stages of collection and storage of metadata and provides evidence for:

- uniquely identifying digital objects;
- certifying the authenticity and integrity of informations;
- documenting the chain of custody of digital objects;
- documenting the technical processes executed for permanent preservation of digital objects;
- informing about the conditions and rights of access to digital objects by final users.

Each metadata format used inside the AP is associated with a namespace, fixing the terminology used, and with a XML Schema which determines its syntactic structure.

The metadata set for MAG is based on the distinction from different types of digital objects (images, OCR texts, sound, audiovisual, digital born text, etc..) rather than from particular types of source documents. The scheme is composed of several sections, whose use, excepting some general areas, depends on the type of digital contents and their use.

The METADIGIT root element contains nine sections:

- GEN: Project infos;
- BIB: descriptive metadata;
- STRU: structural metadata;
- IMG: metadata for still images;
- OCR: metadata for OCR text;
- DOC: metadata for digital objects in text format, derived or digital born;
- AUDIO: metadata for audio files;
- VIDEO: metadata for video files;
- DIS: metadata for the distribution of digital objects.

About the relations between this Italian schema and internationally accepted standards, like METS, it's important to remind that MAG was conceived to collect management metadata about digital objects produced inside a cultural heritage digitalization project. METS is for sure a powerful "packaging schema" with no direct solutions to the requirements – specific and limited – on which MAG is based.

In particular, if we take a look to the technical metadata about digitalization of images – <img> section of MAG – the international work was still in progress and the NISO MIX standard [17] was just in a draft status. Moreover, the audio and video digitalization there were still few referential experiences. Anyway, it has to be considered the implementations registry of METS [18] where for instance someone used – correctly - MAG as an "extension" of METS.

In this direction, the MAG Committee is presently developing the mapping references MAG-METS and MAG-MPEG21-DIDL. MAG takes in both cases the role of a sub-set of a METS or a MPEG21-DIDL metadata document.

In addition, a MAG-PREMIS crosswalk study was started to ensure a correct implementation of MAG-based digital archives that consider the PREMIS model, guaranteeing their long-term management and preservation.

The area of most immediate application of MAG was given by the projects destined to be published in the *Internet culturale* portal [19], that offers didactic, professional and institutional information concerning the Italian cultural heritage and related activities.

In this phase of dissemination a misuse of the MAG application profile has to be quoted: the <bib> section, containing descriptive metadata on digital objects, was often used as a substitute of (digital) catalog descriptions, when missing. The result is that the retrieval system is based on synthetic and Dublin Core -based descriptive metadata, with an item-based granularity (so losing the original collection-level informations) and with an often incorrect interpretation of single semantic elements. The principle to be reminded is that descriptive metadata should not substitute a catalog, but they are useful to retrieve digital objects and to manage the structural and management issues related to digital collections.

Another interesting application of MAG in Italy was for the project SIAS – *Sistema Informativo degli Archivi di Stato*, a national information system, started in 2003, concerning the documentary heritage of the 100 and more Italian State Archives [20].

Inside this application scenario, the MAG metadata AP was adopted more correctly for almost two among its aims:

- to manage the digital repository of digital reproductions of archival documents coing form different institutions towards the national repository and the central web services,
- to ensure a stable link between the reproductions (digital objects) and the archival descriptions included in SIAS digital finding aids, considered as a requirement for the financing of digitalization projects.

## 5. CONCLUSION

Taking for granted that cultural heritage digital applications create digital libraries, every project have to face the challenge of choosing a framework of metadata to guarantee the sound management of its life-cycle, form creation to preservation passing through data delivery. The right choice of on or more metadata application profiles depends both on the current state of the art of metadata standards and on each specific scenario of application. Some misunderstandings for example have been done in applying administrative schemes with the goal of building retrieval base for digital collections.

Thus, a closer exchange of experiences (good practices and typical critical issues) have to be promoted also in the digital experts community.

## REFERENCES


[1] L.Candela et al., *The DELOS Digital Library Reference Model.* Version 0.98. DELOS, December 2007. Available in http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf.

[2] AIB, Gruppo di studio sulle biblioteche digitali, *The Digital Libraries Manifesto*. English version. 2005. Available in http://www.aib.it/aib/cg/gbdigd05a-e.htm3.

[3] Research Councils UK (2008). *Code of Conduct and Policy on the Governance of Good Research Conduct: Integrity, Clarity, and Good Management.* Public Consultation Document. July – October 2008. Available in http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/reviews/grc/consultation.pdf .

[4] CCSDS (Consultative Committee for Space Data Systems) (2002). R*eference Model for an Open Archival Information System (OAIS)*. Blue Book, Issue 1. Washington, DC (US): CCSDS Secretariat, January 2002. Technical report. CCSDS 650.0-B-1. Recommendation for Space Data System Standards. Available in http://public.ccsds.org/publications/archive/650x0b1.pdf.

[5] *MINERVA Technical Guidelines for Digital Cultural Content Creation Programmes*: Version 2.0, 2008. Editors: Kate Fernie, Giuliana De Francesco and David Dawson. Available in http://www.minervaeurope.org/interoperability/technicalguidelines.htm.

[6] The University of Melbourne, Metadata @ Melbourne. *Types of Metadata*. 24 July 2006. Available in http://www.infodiv.unimelb.edu.au/metadata/add_info.html.

[7] R. Wendler, *LDI Update: Metadata in the Library*, in: “Library Notes”, n. 1286 (1999), pp. 4-5.

[8] NISO. *Understanding Metadata*. NISO Press. Avaliable in http://www.niso.org/publications/press/UnderstandingMetadata.pdf.

[9] Bretherton, F. P.; Singley, P.T. (1994). *Metadata: A User's View*, Proceedings of the International Conference on Very Large Data Bases (VLDB). pp. 1091–1094.

[10] LOC, *Standards at the Library of Congress.* Available in http://www.loc.gov/standards/.

[11] Priscilla Caplan, *Understanding PREMIS.* February 1, 2009. Available in http://www.loc.gov/standards/premis/understanding-premis.pdf.

[13] Ministero per i beni e le attività culturali, *Biblioteca Digitale italana. Obiettivi e Contesto*. Available in http://www.bibliotecadigitaleitaliana.it/genera.jsp?s=95&l=it.

[14] ICCU. *Standard MAG - Versione 2.0.1*. Available in http://www.iccu.sbn.it/genera.jsp?id=267.

[15] *MAG: Metadati Amministrativi Gestionali [Administrative Metadata Management] Committee*, http://www.iccu.sbn.it/genera.jsp?id=99&l=en.

[16] MAG 2.0.2. *Reference.* English version, edited by P. Feliciati. 2009. Available in http://www.iccu.sbn.it/upload/documenti/MAG-Reference-201-en.pdf?l=it.

[17] Library of Congress - NISO Technical Metadata for Digital Still Images Standards Committee, *NISO Metadata for Images in XML (NISO MIX). 2.0.* Available in http://www.loc.gov/standards/mix/.

[18] http://sunsite.berkeley.edu/mets/registry/.

[19] http://www.internetculturale.it/genera.jsp?s=1&l=en.

[20] Ministero per i beni e le attività culturali, *Sistema Informativo degli Archivi di Stato*. Available in http://www.archivi-sias.it .

[21] P. FELICIATI, (2007). *Dalla descrizione archivistica al documento digitale: l'adozione del profilo MAG per la gestione della digitalizzazione negli archivi storici.* Digitalia, vol. 1; p. 35-48, available in http://digitalia.sbn.it/.