The bibliographic information of electronic documents

Catherine Lupovici,
Library activities Manager

Abstract

The bibliographic information is value added information to primary documents created to facilitate information retrieval and document access. The extension of the classical Library catalogues, data bases and archival findings aids functions to the electronic documents handling is to provide a link to the document itself. For this purpose the MARC formats are introducing a specific field to handle the linking information. At the same time other approaches are tested by different users communities to add the bibliographic value added information into the electronic document itself to enhance the search and access functionality of the documents data bases. This approach uses the electronic document format itself. Such are the Dublin Core metadata approach based on HTML, the Text Encoding initiative (TEI) approach for electronic documents in humanities and linguistic arena based on SGML. TEI is also applied in the cultural heritage document conversion domain with the creation of DTDs for the single document or for a collection of documents. All this projects and approaches are contributing to build the future of digital libraries and archives.

Introduction

The bibliographic information is created to facilitate the selection of and the access to the primary document. This value added information is encoded in a format allowing the preparation of different products which are currently printed bibliographies, online services and CD-ROM products. This secondary information is composed mainly of two parts : the descriptive part and the access points. The descriptive part is a compilation of data taken from the document itself such as title, responsibility statement, publishing information, physical format and sometimes of additional comments written by the cataloger according to standard rules. The access points are the true value added information. They are generally compliant with thesaurus or authority forms for names, corporate names, key-words.

Automated bibliographic databases are, since 30 years, handled through a bibliographic record and the access points are controlled by external dictionaries of allowed forms. The link with the described document is provided through a call number giving the location of the physical item.

Handling the electronic document needs to extend the existing organization or to rethink it, depending of the relative amount of items of each type and depending of the services offered. The simple extension of capabilities to handle the electronic document in the same way as the classical one is to provide the link from the bibliographic record towards the document itself for access. In that case, one may think to extend the concept of value added data from the classical headers or key-words to full other documents providing added value to the original one.

We can also consider the electronic documents repository that can offer direct access to the full document, depending of the technical capabilities offered by the encoding format used. In that case the descriptive data and sometimes more can be found directly in the document itself and the value added access points can be added directly to the document in the document format.

All this approaches are experimented in the digital library or digital archives environment and mainly in the context of digitalization programs for which libraries or archives can choose the collection, the format of the electronic items as well as the whole system functional architecture.

1. Extension of the bibliographic format

The extension of the classical document processing philosophy to the electronic document covers two main points

- The creation of the descriptive and access points data. For pure electronic documents it is sometimes very much time consuming to extract manually the data from the document which needs to be installed and de-installed. For reproduction in electronic form of existing classical documents it can be just adding the reproduction information to existing data
- The creation of a kind of call number to allow the user to consult the document. This information can be proprietary or can use de facto standards for location such as the URL (Uniform Resource Locators) in a Web environment.

Formats for bibliographic data such as the MARC formats very much in use in the Library community can be used for this purposes, allowing to give the reproduction information in a specific field.
In addition the 856 field has been created in USMARC and UNIMARC for the electronic information and access. In UNIMARC [i] this field is defined for the entire title of a publication (the journal title for instance and not the issue). It is a repeatable field for different locations or different file names. The following indicators and subfields are defined :

Indicator 1: Access Method (email, ftp, telnet, dial-up, http, specific)
Indicator 2: Blank (not defined)
Sub-fields

$a Host name
$b Access number (IP address, telephone number)
$c Compression information
$d Path
$f Electronic name (of file or files)
$g Uniform Resource Name (URN)
$h Processor of request
$i Instruction (if needed by a remote host to process a request)
$j Bits per second
$k Password (general–use passwords, not security passwords)
$l Logon/login (general-use logon/login)
$m Contact for access assistance
$n Name of location of host in sub-field $a
$o Operating system of host in sub-field $a
$p Port
$q Electronic Format Type (ASCII, MIME Internet media types)
$r Settings used for transferring data
$s File size
$t Terminal emulation
$u Uniform Resource Locator (URL)
$v Hours access method available
$w Record control number
$x Nonpublic note
$y Access method (if not one of the three main TCP/IP protocols)
$z Public note

Examples :

856 3#$b1-202-7072316$j2400-9600$nLibrary of Congress, Washington, DC$oUNIX$rE-7-1 $tvt100$zRequires logon and password

200 0#$aBulletin d'informations de l'Association des bibliothécaires français
856 4#$uhttp://www.abf.asso.fr/bulletin.htm$zSommaire des numéros disponible en ligne

2. The Dublin Core metadata approach

In the Web environment one can use the metadata feature to embed in a Web HTML page a description on the page resource using a specific syntax. Such data are used automatically by web crawlers.
Within this general metadata Web framework the Dublin Core[ii] initiative started an international standardization work to define data elements for bibliographic information to be embedded in the Web pages. The data elements list was established in December 1996 and they are under experimentation in different projects with sometimes different interpretations in different implementations.
The Dublin Core (DC) metadata are a consensus on a minimal electronic resource description that can be used from the creation to the search and retrieve processes in the Web environment.
It can also be used as a minimum set of elements for interoperability between more sophisticated formats.

The data elements defined in DC are :
Content metadata :

- Title
- Subject : topic of the resource
- Description : a textual description of the content of the resource
- Source : resource from which this resource is derived
- Language
- Relation : relationship to other resources
- Coverage : spatial or temporal characteristics of the intellectual content of the resource

Intellectual property metadata :

- Creator : primary responsibility of the intellectual content
- Publisher : entity responsible of making the resource available in the present form
- Contributor : person or organization who have made a significant intellectual contribution to the work
- Rights : link to right management statement or service giving such information

Instantiation metadata

- Date
- Type : category of resource e.g. home page, poem, working paper
- Format : data format (software and hardware needed to exploit the resource)
- Identifier : string or number used to uniquely identify the resource (URL, URN, ISBN ...)

A controlled vocabulary can be associated to the content of some data elements. The DC data elements are experimented in a lot of projects in different countries as well as in European Library projects. Some significant examples related to bibliographic information in European countries are :

Germany : Metadaten-Projekt = Metadata Project. It explores the use of metadata from a library point of view and looks at the impact of the developments in networked information resource discovery on traditional cataloging rules.
Home Page: http://www2.sub.uni-goettingen.de

Scandinavia :
The Nordic Metadata Project. A share metadata creation system. DC is being used to make available to the end-user a diversity of digital documents over the Net
Home Page: http://linnea.helsinki.fi/meta/

INDOREG: INternet DOcument REGistration. A project of the Danish Library Center (DBC) to provide registration of all internet publications which fall into specific inclusion criteria and to provide access to these documents through DanBib.
Home Page: http://www.purl.dk/rapport/html.uk/

United Kingdom :
Project BIBLINK. European Library project of several national libraries which aims to establish an electronic metadata link between publishers and National Bibliographic Agencies (NBA's) to exchange metadata records of newly published items.
Home Page: http://www.ukoln.ac.uk/metadata/BIBLINK/

Electronic Library Image Service for Europe (ELISE II). In the ELISE II prototype the catalogue data supplied by participating institutions is mapped to DC and displayed alongside thumbnail images.
Home Page: http://severn.dmu.ac.uk/elise/

3. SGML and bibliographic information

SGML (Standard Generalized Markup Language), ISO 8879 international standard is being used in several digital resource projects where the document and the bibliographic information are handled with the same high level professional format.
SGML technology is used in two ways to handle bibliographic information and electronic documents at the same time.
One approach is to add an SGML header to the structured document applying the TEI framework and giving the bibliographic information item by item at the lower level of granularity. The second approach is to define an SGML DTD enabling to handle the whole collection information through a hierarchical information going from the collection description root to the piece level, through the collection organization. It reproduces the hierarchy of archival finding aids tools like inventories, registers, indexes, and guides.

3.1 Structured text headers approach

The research community interested in computer use in Humanities, Literature and Linguistic started ten years ago to develop on SGML a common encoding scheme for use both in creating new documents and in exchanging existing documents among text and data archives. This project known as TEI[iii] (Text Encoding Initiative) resulted into an SGML DTD complemented by Guidelines for Text Encoding and Interchange. A TEI Lite DTD was also created with the basic set of elements that can be use for general text types.

One of the features of the TEI DTD is to define a document header providing information on the marked up text such as the source, the markup principles, information on history of the text revisions

and modifications. The header tag is mandatory in a TEI document. The TEI header is composed of the four following parts :

- File description (mandatory) : it is an electronic equivalent of a title page for a paper document. The flexibility of the TEI framework allows for instance to describe a text according to the AACR2 (Anglo American Cataloging Rules)
- Encoding description : relationship between the encoded text and its source(s). It can be for instance the encoding project name or the editing decisions taken
- Revision description : contains the history of revisions

The header can be very simple or complex, depending of the application needs.

Several specific TEI compliant DTDs were developed and are currently experimented in electronic resources projects. Two significant projects are :

Electronic Text Center, University of Virginia[iv]**.** This center, established in 1992, offers an on-line archive of thousands of SGML encoded electronic texts. They are using the TEILITE.DTD for document encoding. The search capability is offered to users through a Web interface and documents are reformatted on-the-fly into HTML for display. In this application the TEI header is very complex : it is a record of the printed source of the electronic text, of the creation of the electronic text and it provides various information for the search tools. It is also the source of the USMARC record that goes into the library catalog. The USMARC record is generated through an automatic conversion program from the header.

Library of Congress American Memory DTD for Historical Documents[v]. In its presentation of historical collections, the Library of Congress uses SGML for marking up the full text of books, pamphlets, manuscripts and other historical texts. The American Memory DTD (AMMEM.DTD) used for this process is a TEI Lite application. The AMMEM.DTD files are available from the LC Web server.
For this application, the Library of Congress decided not to replicate the full bibliographic information in the header. Only the title from the USMARC 245 field is copied and complemented by the phrase "a machine-readable transcription". The LCCN (Library of Congress Card Number) is also provided within the publication statement, if such a number has been assigned for the bibliographic record of the source document.

3.2. The Encoded Archival Description (EAD) DTD

Encoding the archival finding aids in SGML began at the University of California, Berkeley in 1993. The first DTD version was then tested by other libraries and enhanced including participation of the Committee on Archival Information Exchange of the Society of American Archivists. The institutions involved in the working group included the Library of Congress, RLG, OCLC and SAA. A Beta test version was issued in July 1996 and is being experimented in several projects. As a potential international standard it is maintained at the Network Development and MARC Standards Office of the Library of Congress in partnership with the Society of American Archivists.
The goal is to provide a non-proprietary encoding standard for machine-readable finding aids together with the desire to go beyond that information provided by traditional MARC records. The EAD.DTD accommodates registers and inventories of any length describing the full range of archival holdings. It covers also textual electronic documents, visual materials, and sound recordings. The header is TEI compliant. The classical descriptive paper reference tool is converted into a potential powerful object-oriented data base.

The standard provides gateways to the traditional formats and tools since it offers MARC equivalency attributes for use with elements matching USMARC fields, including the possibility to add the authoritative MARC form. The elements defined in the EAD.DTD are compliant with the ISAD(G) (General International Standard Archival Description)

Some examples of the EAD DTD implementations are :

- **California Heritage Digital Image Access Project.** The project involves embedding digital representations of original archival materials within finding aids that have been encoded with the beta EAD.DTD. The finding aids are displayed in HTML through an on-the-fly SGML conversion . Home page : **http://sunsite.berkeley.edu/CalHeritage**

- **Library of Congress Finding Aids Project.** The home page of the project provides access to all the archival findings aids at the LC which have been encoded using the EAD DTD. Future improvements will offer to search across encoded finding aids for all the LC divisions and links from the catalog bibliographic records to finding aids. Home page : **http://www.loc.gov/rr/ead/eadhome.html**

- **American Heritage Virtual Archive Project.** Collaborative project between four universities to explore the various factors involved in creating and maintaining a share union data base of EAD finding aids associated with digital representation of items described in such finding aids. Home page : **http://sunsite.berkeley.edu/amher**

Conclusion

The different formats for creating and handling the bibliographic information which are under development are of course offering different functional architecture capabilities. They are also representing different possible steps of migration towards the new electronic environments.
The simple extension of the bibliographic information with a link towards the document reproduction provides the classical access to the electronic document which can be manipulated separately in the technical context of the separate platform or platforms supporting the documents. There is no fundamental technical changes for the institution. The documents are processed in the classical chain with the same philosophy as the other materials. Handling the electronic documents is a separate new task with no direct impact on the existing library or archive systems and organization.
A step forward is to manage electronic documents in HTML and to create Dublin Core metadata in the corresponding Web pages. The documents are directly searchable in the Web service through the metadata and the full text document. In addition the institution can process the DC information to create a minimum classical bibliographic record loaded in the automated catalog. This minimum level record can be further enhanced if there is resources available in the institution for this task. Such an organization is introducing the philosophy of cataloging in the document itself which is a new concept.
The more elegant and interesting approach is the third one using SGML and particularly where entire collections of objects can be described in a powerful and consistent single model. It is important when there is a need to carry out the description from the collection level to the item level reproducing the whole collection organization and offering at the leave hierarchical level the electronic object itself. Of course we already have the linking facilities in UNIMARC, or the segments facilities in the CCF (Common Communication format), allowing to provide a cascading information. But it restricts us to the close world of library systems. Using SGML opens to build information services on tools and concepts facilitating exchange and share with other communities like museums, researchers, publishers, corporate information including gray literature, technical documentation. It is participating to the on going evolution and learning (the light side) but also

spending time for investing into new organization and spending money into new systems (the dark side). Anyway if you already spent time and money into professional systems based on standards like UNIMARC, you will be able at any time to convert all what you got, and only what you got, into a finding aids and document handling SGML system.

Notes

[1]856 Electronic Location and Access . URL : http://ifla.inist.fr/VI/3/p1996-1/856.htm

[1]Dublin Core metadata. URL : http://purl.oclc.org/metadata/dublin_core/

[1] Text Encoding Initiative. URL : http://www-tei.uic.edu/orgs/tei/index.html

[1] University of Virginia Library. Electronic Text Center. URL : http://etext.lib.virginia.edu/

[1] American Memory DTD for Historical Documents. URL : http://lcweb2.loc.gov/ammem/amdtd.html