# Digital Libraries: An overview of Standards, Protocols and Formats

**Jayant Deshpande**
Librarian,
Indian Institute of Carpet Technology
Bhadohi-221401(UP)

***Abstract*** *- The use of Information and Communication Technology (ICT) & digitization of the libraries facilitate easy & immediate access to information. Digitization of libraries has availed the libraries to keep pace with the latest development .This has additionally facilitated precision, flexibility and reliability in the library and information centre. Digitization of libraries reduces the perpetual work and preserves time and brings precision and speed & keeps material safe. It withal increases efficiency in technical processing of library materials and amend the efficiency of library administration and management. There are numerous challenges in organising a digital library (DL) and successfully get to the user the articles he/she wants. Some of the problems have already been addressed and a few more are yet to be solved. Various standards have already been developed in storage and retrieval of digital data which are described here. They range from the standards that cover Portable Document Format files through the standards that govern international cataloguing efforts to standards for searching. Others are still under development and are described in the state in which they are currently. In conclusion, standards are necessary for every aspect of the digital library. New standards are being developed by the formal international and national standards bodies and one set up especially for the purpose such as COUNTER has been highlighted*

**Key words:** Standards, formats, protocols, retrieving information, ASCII, control information coordination.

## Introduction

The term electronic library, digital library and virtual library have been used interchangeably and now widely accepted as description of the use of digital technology by libraries to acquires, store, conserve to remote users.

"Digitization" refers to all of the steps involved in the process of making collections of historical materials available online. It is the process that creates a digital image form an analogue image".

Digitization refers to the process of translating a piece of information such as a book, journals, articles, sound recording, pictures, audio tapes or video recording etc. into bits.
"Digital library is the concept of information stored digitally and made accessible to users through systems and networks".

"Digital library is essentially a fully automated information system with all resources in digital form".

"Digitization is the process of converting the content of physical media (e.g. periodical, articles, books, manuscripts, cards, photographs, vinyl disks, etc.) into digital format".

The American Digital Library Federation has defined the digital library as "Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collection of digital works so that they are readily and economically available for use by a defined community or set of communities."

**Digital Library, Standards, Protocols and Formats**

There is a great need for adopting various standards and best practices to build interoperable digital libraries. Standards, Protocols and Formats are the rules by which objectives are described, their data is stored and the systems communicate. This is an electric mix of such rules. Some standards are international standards set by bodies like ISO (International standards Organisation) and IEFT (Internet Engineering Task Force). Some are national standards set by bodies like NISO (National Information Standards Organisation) in the U.S. or BSA (British Standards Authority) in the U.K. Some are industry standards set by industry bodies. Some are corporate standards produced by a single company and accepted by widespread usage.

**Types**
Some important standards, protocols and formats which are useful to build digital libraries are listed below:
1) Bibliographic 2) Record structure 3) Encoding 4)Communication 5)Protocols 6) Formats

**PROTOCOLS**

A protocol is defined as a set of rules or conventions formulated to control the exchange of data between two entities desiring a connection. Protocols are required to define the exchange of control information between user device and the network. Basic elements of a protocol include data format and signal levels, control information coordination and error handling, and timing.

**Important protocols:**

- IP- (Internet Protocol)
- TCP (Transmission control protocol)
- FTP (File Transfer protocol)
- Z39.50
- SIP2 (Data Transfer protocol for RFID devices)
- HTTP (Hypertext Transfer protocol)
- OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)

**Simple Dublin Core**
The Simple Dublin Core Metadata Element Set (DCMES) consists of 15 metadata elements:

| | | |
|---|---|---|
| 1. Title | 2. Date | 3. Relation |
| 4. Creator | 5. Type | 6. Coverage |
| 7. Subject | 8. Format | 9. Rights |

| | | |
|---|---|---|
| 10. Description | 11. Identifier | 12. |
| 13. Publisher | 14. Source | 15. |
| 16. Contributor | 17. Language | 18. |

**Qualified Dublin Core** includes three additional elements (Audience, Provenance and Rights Holder), as well as a group of element refinements (also called qualifiers) that refine the semantics of the elements in ways that may be useful in resource discovery.

**BIB-1**

- The Bib -1 is a simplified record structure for online transmission.
- The Bib-1 attribute set is part of the Z39.50 client sever protocol. This set defines six attributes to be used in searches of information on the server computer: use, relation, position, structure, truncation, completeness.

The syntax of the Z39.50 protocol allows for very complex queries.

**Text Encoding Initiative**

The **Text Encoding Initiative (TEI)** is a consortium of institutions and research projects which collectively maintains and develops a standard for the representation of texts in digital form.

**Electronic Archive Description**

An encoding scheme devised within the Standard Generalized Mark-up Language (SGML) framework to define the content designation of document and other archival objects. It is defined with a minimum number of descriptive elements, but in an extensible fashion. It is designed to create descriptive records which will assist in searching for the original material in a number of ways.

**Federal geographic data committee**

A metadata standard for the description of the elements of maps and other cartographic objects, including such attributes as scale, projection, co-ordinates (and co-ordinate scheme), etc. This is just an example of a number of specialist descriptive schemes for different objects and material types.

**Metadata**

- Metadata is data about data. An Item of metadata may describe an individual datum, or content item, or a collection of data including multiple content items.
- Metadata (sometimes written 'meta data') is used to facilitate the understanding, use and management of data. The metadata required for effective data management varies with the types of data and context of use. In a library, where the data is the content of the titles stocked, metadata about title would typically include a description of the content, the author, the publication date and the physical location. In the context of a camera , where the data is the photographic image, metadata would typically include the date the photograph was taken and details of the camera settings.

**Anglo-American Cataloguing Rules (AACR)**

These are a set of cataloguing rule which define how and object is to be described. They are entirely intellectual and are concerned with such things as the consistency of people's names and subject descriptions. They are not absolutely tied to any standard format, though they are

developed in conjunction with MARC records. They (or a close derivative) are in use in all types of libraries around the world. Russian and German libraries have developed their own similar rules for the same purpose (GOST and RAK), but AACR2 is by far the most common code of cataloguing.

**Classification Schemes (Dewey, UDC, BSO, and etc.)**

These are intellectual schemes for the ordering of knowledge and are used for assigning a work to a class along with works of similar content. They are also used as the basis for the physical ordering of physical objects (books, tapes, etc.) on shelves. They could be used for assigning a 'location' to an electronic object – most likely as a way of deriving a unique file name for the object. Dewey and UDC schemes are far and away the most common schemes of classification.

**URLs**

These are Uniform (or Universal) Resource Locators and are the addresses of objects within the Internet. As such they satisfy the requirement for Uniform, but because they are limited to Internet (and more generally World Wide Web) use they are not Universal. They are the links between sites or pages on the web, which allow the linking (or Hyper linking) and provides the navigational functionality of the web. They are not bibliographic in nature, but can be used to provide a non-liner logical structure to documents on the Internet's.

**Record Structures**

Record structure defines the physical and logical structure of the record which holds the data. The very simplest of them hold only a single type of data (such as an image) and they are listed later in the section on formats. The records considered here are complex in that they contain

Multiple fields of variable length and which may occur more than once. Except for proprietary structures there is really only one structure used for bibliographic data of any complexity. These formats are for exchanges of data between systems and are not intended for human consumption.

**ISO 2709**

ISO 2709 is an ISO standard for bibliographic descriptions, entitled Format for bibliographic Information Interchange on Magnetic Tape. A ISO 2709 record has four sections:

- Record label
- Directory
- Data fields
- Record separator

**2.2 Z39.2**

This standard specifies the requirements for a generalized interchange format that will accommodate data describing all forms of material. It describes a generalized structure designed specifically for exchange of data between processing systems and not necessarily for use as a processing format within systems. It may be used for the communication of records in any media.

**Encoding**
Encoding is the process of transforming information form one format into another. The opposite operation is called decoding. Theoretically, other material objects (such as images or sounds) could be encoded within a complex record (where there could be a text field, an image field and a sound field), but it is very rarely done. These combinations are done either in a meta-record (such as the bibliographic record) or within the text record itself. The other records are referenced (as in an HTML page with a reference to an image file).

**Unicode**
Unicode is an industry standard allowing computers to consistently represent and manipulate text expressed in most of the world's writing systems. This is a universal encoding scheme using 16 bits to represent each character. It has the advantages of being simple, complete, and is being widely adopted. Its disadvantage is that all characters take twice as much space even for single language data. Unicode is controlled by the Unicode consortium and is the operational equivalent of the ISO-10646 standard. Note that 10646 also define 32bit characters, but these are not in any general use.

**ASCII**
ASCII specifies a correspondence between digital bit patterns and the glyphs (i.e., symbols) of a written language. This allows digital devices to communicate with each other and to process, store, and communicate character-oriented information. The ASCII character encoding-or a compatible extension (see below)- is used on nearly all common computers, especially personal computer and workstations. The preferred MIME name for this encoding is "USASCII".

**Communications Formats**
There are many layers of communications (seven if one considers the 'OSI' seven layer model), most of which do not concern us here. The one that is important is the level at which the computer systems connect to each other to create a connection that will pass our messages back and forth. There is one protocol for this level that is by far the most common. It is also the protocol of the Internet.

**TCP/ IP**
This protocol called Transmission Control Protocol / Internet Protocol (TCP/IP) is for controlling the creation of transmission paths between computers on a single network and of connecting between different a single network and of connecting between different networks. It is in almost universal use for public networks and many in house local area networks. It is the protocol of choice for all UNIX servers (Sun uses it universally) and most workstations. The only reason not to specify it is if your systems will be entirely in- house and the existing networks uses something else.

**Protocols**
These are the 'language' of the messages passing between the systems connected via a TCP/IP (or other) protocol. There are a variety of protocols for different purpose, which may be used at different times by the same two systems or by one system 'talking' to two others.

**http**
Hypertext Transfer Protocol (HTTP) is a communications protocol used to transfer or convey information on intranets and the World Wide Web. Its original purpose was to provide a way to publish and retrieve hypertext pages, Development of HTTP was coordinated by the W3C

(World Wide Web Consortium) and the IETF (Internet Engineering Task Force), culminating in the publication of a series of RFCs, most notably RFC 2616(June 1999), which defines HTTP/1.1, the version of HTTP is common use.

**ftp**

**FTP** or **File Transfer Protocol** is used to transfer data from one computer to another over the Internet, or through a network. Specifically, FTP is a commonly used protocol for exchanging files over any network that supports the TCP/IP Protocol (such as the Internet or an intranet).

**Z39.50 ISO 23950 (ISO-10162/3)**

Z39.50 is a client server protocol for searching and retrieving information from remote computer databases. It is covered by ANSI/NISO standard Z39.50 and ISO standard 23950. The standard's maintenance agency is the Library of Congress. Z39.50 is widely used in library environments and is often incorporated into integrated library systems and personal bibliographic reference software. Interlibrary catalogue searches for interlibrary loan are often implemented with Z39.50 queries.

**Z.39.63 (ISO-10160/1)**

These are the NISO and ISO interlibrary loans protocols. Unlike the Z39.50 (ISO10162/3) Protocols they are not identical, merely functionally equivalent. Most ILSs support these functions through a standalone module and some through access to third party services. The ILL protocol allows various options for implementation, such as the set of messages to be supported and the message transport mechanism. An example of a third party ILL service is: http://www.cps-us.com 2.6 formats.

**File Formats**

Image file formats provide a standardized method of organizing and storing image data. This article deals with digital image formats used to store photographic and other image information. Inage files are made up of either pixel or vector (geometric) data. Which is rasterized to pixels in the display process, with a few exceptions in vector graphic display? The pixels that comprise an image are in the form of a grid of columns and rows. Each of the pixels in an image are in the form of a grid of columns and rows. Each of the pixels in an image stores digital numbers representing brightness and colour.

**Image**

**JPEG; BMP; TIF; GIF; PNG; RAW; SVG**

**Animation**

- ANI
- FLI
- FLC
- 6.3 Video
- AVI
- MOV
- MPG
- QT
- 6.4 Audio
- WAV
- MID
- SND
- AUD
- 6.5Web Pages
- HTM
- HTML
- DHTML
- HTMLS
- XML
- 6.6 Text
- DOC
- TXT
- RTF
- PDF
- 6.7 Programs
- COM
- EXE

## Conclusion

When creating digital library systems which contain valuable contain valuable content, we are making important promises to both current and future users. For this, one cannot forget the good standards, Standards play most important role for the effective utilization of networked information as well as the development of digital libraries. While libraries have been using standards for bibliographic description and interchange of cataloguing data for many years, they need to develop a new level of standards awareness to handle the networked information environment. The new standards must address information technology and communication, information search and retrieval, text encoding, open system interconnection, mark-up languages, electronic document interchange and ILL protocols.

## References:

1. Application Service Definition and Protocol Specification. Washington DC, NISO, 1995 (Z39.50)

2. CCF/B : the Common Communication Format for Bibliographic Information. Paris, UNESCO, 1992.

3. Common Command Language for Online Interactive Information Retrieval, Washington, DC, 2008 (Z39.58)

4. Counter –Counting Online Usage of NeTworked Electronic Resources http:// www.projectcounter.org/

5. Dierickx, H. and Hopkinson, A. Reference Manual for machine-readable bibliographic descriptions. 2nd ed. Paris, UNESCO, 1981

6. Digital Object Identifier (DOI). Geneva, ISO, to be published. (ISO/DIS 26324)

7. Document management — Portable document format — Part 1: PDF 1.7 PDF (ISO 32000-1:2008)

8. Document management — Portable document format — Part 1: PDF 1.7. Geneva, ISO, 2008 (ISO 32000-1:2008)

9. Format for Information Interchange. Geneva, ISO, 2008 (ISO 2709:2008)

10. Holdings statements for bibliographic items. Washington DC, NISO, 2006. (ANSI Z39.71)

11. Information and documentation — Commands for interactive text searching. Geneva, ISO, 1993 (ISO 8777:1993)

12. Information and Documentation. Information Retrieval (Z39.50). Geneva, ISO, 1995 (ISO 23950)

13. Information and Documentation: Holdings Statements: Summary level. Geneva, ISO, 2006 (ISO 10324:1997)

14. Information interchange format. Washington, DC, NISO, 2001. (Z39.2 )

15. Joint Steering Committee for Revision of AACR .Anglo-American Cataloguing Rules 2nd ed., 2002 Revision: 2005 Update. Washington DC, ALA; Ottawa, CLA; London, CILI, 2005.

16. MARC21 Format for Holdings Data. 2000 ed., update 9. Washington DC, Library of Congress, 2008 http://www.loc.gov/marc/ holdings/echdhome.html

17. MARC21 is found at http://www.loc.gov/ marc/- 105 7th International CALIBER 2009 Challenges for the Digital Libraries and Standards to Solve ...

18. Metadata Object Descriptions Schema (MODS). Version 3. Washington DC, Library of Congress, 2008

19. NISO Circulation Interchange Part 1: Protocol (NCIP). Washington DC, NISO, 2008 (Z39.83-1). NISO Circulation Interchange Protocol (NCIP) Part 2: Implementation Profile 1. Washington DC, NISO, 2008 (Z39.83-2).

20. ONIX for Serials: SOH: Serials Online Holdings -Version 1.0 (revised September 2005), Version 1.1 (June 2007). Washington DC, Editeur, 2007

21. Standardized Usage Statistics Harvesting Initiative (SUSHI) Protocol. Washington DC, NISO, 2007. (Z39.93)