Interoperability among Multi-Lingual Digital Libraries through UNICODE based metadata: a Model for India

Position Paper

Indo-US Workshop on Open Digital Libraries and Inter Operability, Virginia Tech, USA
23-25 June 2003

Dr T A V Murthy,

Director, **INFLIBNET**/UGC, Ahmedabad-380009, India | tav@inflibnet.ac.in | www.inflibnet.ac.in

Abstract

Digital Libraries represent a new infrastructure and environment that has been created by the marriage of computing, communication and content, on a global scale. This supports individuals or organizations in a good broad range of distributed knowledge based activities from electronic commerce to scientific collaboration. It is essential that this new type of functionality is developed in order to allow us to solve complex global challenges in different areas. One of the areas of libraries in this global scale is e-content and knowledge explosion. A collaborative effort among leading researchers from the US and India could explore the possibilities of a joint international research agenda in the field of Open Digital Libraries with sufficient emphasis on Interoperability, Global resource Discovery, Metadata, and Multilingual Information Access etc. Out of these areas, this paper shares the ideas relating to the interoperability and multilingual issues of the Indian Digital libraries. India with its rich diversity of languages and commonality can safely address UNICODE-metadata as an interface to represent the multilingual issues.

Interoperability

Interoperability is the ability of digital library components and services to be functionally and logically interchangeable by virtue of their having been implemented in accordance with a set of well-defined publicly known interfaces.

Tools

Interoperability is a critical problem in the network environment especially when we are talking about the Digital Libraries with increase in number of diverse computer systems, software applications, file formats, information resources and users. But it becomes more critical problem in Indian digital libraries, with having those much differences it has another sharing problem of resources from one language to another as resources at Indian libraries are present in many Indian languages viz. English, Hindi, Sanskrit, Marathi, Gujarati, Oriya, Bengali, Punjabi etc. Thus it has

problem of interoperability between multilingual digital library resources. However there are so many true type fonts are being used to represent the Indian languages on web. But that's not sufficient tool to implement the multilingual. ISCII is also being used as a standard to represent the Indian languages on the web as well on the database part. But as we found that Unicode is the only solution, which represent the all languages being spoken in this world including the Indian languages. It does not have any ambiguity and doesn't overlap, dealt with each and every characters of every language with unique values. To implement effectively interoperability function within the digital libraries following standards speak as tools:

Metadata

In simple way "Metadata" is a data about data, but basically it is a "structured data about data". It gives information about the data, which are stored on web. Each and every page of any website concerns with the metadata. Those metadata keeps information about the page on which the page talks. Different metadata standards are being used to represent the records in web such as MARC, Dublin Core, BIB-1, Text Encoding Initiative (TEI), and Electronic Archive Description (EAD) etc. Dublin Core is defined to be simple enough for people unschooled in the science of cataloguing to tag their documents for indexing by web harvesters while MARC has a richer description formats. Interoperability between digital libraries needs standardization in metadata.

UNICODE

Unicode is an encoding scheme, provides a unique number for every character, no matter what the platform, no matter what the program, and no matter what the language is. It is not a hardware or software, it is a formal standard. It enables a single software product or a single website to be targeted across multiple platforms, languages and countries without re-engineering. As per the definition of the multilingual digital library "A multilingual digital library is a digital library that has all functions implemented simultaneously in as many languages as desired and whose search & retrieve functions are language dependent". Thus the slogan given for the Unicode by the consortia becomes true, i.e. "When the world wants to talk, it speaks Unicode". It allows data to be transported through many different systems without corruption due to its platform independent capability.

Chapter 9 of the Unicode standard book names as "South and South-East Asian Scripts" covers almost 15 different scripts. Out of these 15, all 18 constitutionally recognized languages (covered by 9 different scripts and allocated the unique values to each every character of the languages. In

addition to these languages, it also supports different Indian dialects such as Awadhi, Bagheli, Bhatneri, Bhili, Bihari, Braj Bhasha, Chhattisgarhi, Garhwali, Gondi (Betul, Chhindwara, and Mandla dialects), Harauti, Ho, Jaipuri, Kachchi, Kanauji, Konkani, Kului, Kumaoni, Kurku, Kurukh, Marwari, Mundari, Newari, Palpa, and Santali etc. Following table represent to allocated unique numbers to the Indian scripts:

Script	**Assigned unique number by Consortium**	
Arabic	U+0600 – U+06FF	(01536 - 01791)
Devnagari	U+0900 – U+097F	(02304 – 02431)
Bengali	U+0980 – U+09FF	(02432 – 02559)
Gurumukhi	U+0A00 – U+0A7F	(02560 – 02687)
Gujarati	U+0A80 – U+0AFF	(02688 – 02815)
Oriya	U+0B00 – U+0B7F	(02816 – 02943)
Tamil	U+0B80 – U+0BFF	(02944 – 03071)
Telugu	U+0C00 – U+0C7F	(03072 – 03199)
Kannada	U+0C80 – U+0CFF	(03200 – 03327)
Malayalam	U+0D00 – U+0D7F	(03328 – 03455)

Technology for Multilingual digital libraries:

- **Unicode based Operating System:** Windows 2000/XP, NT, AIX, Sun Solaris, HP/UX
- **Unicode based Front-end Software:** HTML, XML, Java, JavaScript, .NET, PERL, XML, ASP, JSP, CORBA, etc
- **Unicode based Back-end Software (RDBMS):** MS SQL Server 2000, Sybase, Oracle 8i and DB2.
- **OCR (Optical Character Recognition) Software**

Sorting Algorithm:

Sorting algorithm is defined as per the following three different techniques:

- *Primary Only* means that only major differences are considered, such as different base letters (e.g. *A vs B*).
- *Primary & Secondary* means that a second level of differences is also considered, such as accents (e.g. *A vs Á*). However, these are only relevant if there are *no* primary differences anywhere else in the strings.

- *Primary - Tertiary* means that a third level of differences is also considered, such as case (e.g. *A vs a*). However, these are only relevant if there are *no* primary differences and no secondary differences anywhere else in the strings.

The above algorithm works for any languages. But another algorithm needs to be defined for the priority of the language to be indexed first in case of sorting of mixture of languages.

Harvesting protocol

The Open Archive Initiative (OAI) for metadata harvesting is a new protocol dedicated to solving problems of digital library interoperability by defining simple protocols, most recently for the exchange of metadata. Although Z39.50 Information Retrieval Protocol is also one the harvesting protocol. Some digital library systems are using for retrieving and searching the records for example Maxwell System. It is crucial to explain, discuss, and disambiguate the concepts and terminology used among OAI implementers to harvesting approach to interoperability.

The harvesting protocol defined by the OAI is a request / response protocol with 6 request types, viz. *Identity, ListMetadataFormats, ListSets, GetRecords ListRecord, and ListIdentifiers.* It has defined in such a way to support a common set of principles and a technical framework to achieve interoperability.

Functionality Specifications:

Functionality wise the multilingual digital library should have flexible with:

- *Comfortable to the user*
  Digital libraries are created in countries and, in general may be somehow specialized. It is expected that the users in the country where the library operates will want to use it in their native language. At the same time, users with other native languages than that of the country under consideration may need more international languages, as for example, English or Hindi

- *Flexible in terms of the chosen languages*
  Digital libraries in different countries and aiming at different sets of users may operate with distinct sets of languages. As an example, in India a good set of languages is Hindi,

Bengali, Gujarati, Marathi, Oriya, Sanskrit, Kannada, Telugu, Tamil and English etc. But in USA or in UK, it must be different

- *Wide and accurate in terms of the information*
  In order to fulfill its functions, the digital library must be accurate in terms of information. This requires that the languages be kept track at any moment of operation by the user; he /she must know the languages of the catalogue entry and of the content, regardless the navigation language in use.

  For the search to be effective, no matter the navigation language, the user may submit search arguments in any language and all points of access must be possible to be searched in all languages, regardless of the navigation language of the session.

  In order to identify contents in original languages and corresponding translations into other languages, a strict translation control must exist.

- *Easy to operate*
  The digital librarian must control all cataloguing (original languages and translations) and all the contents (original languages and translations). For data integrity to be achieved, translation interfaces for the cataloguing and translation control applications must be available.

Language control Parameters:

- *Language of the content*: it is one of the metadata in metadata schemes
  The language of the content is the language in which the content is written and/or spoken
- *Language of the catalogue entry*
  The language of the catalogue is the language in which the attributes of the content and its instances are written

- *Language of navigation*
  The language of navigation is the language of all the interfaces and messages of the digital library system

Multi Lingual Digital Library Model for India