15. Localisation & Language Technology Standard : An overview

Dr. Om Vikas, Mr. Manoj Jain, Mr. Vijay Kumar, E-mail : tdilinfo@mit.gov.in

Abstract:

This paper presents a quick overview of the current Language Technology & Localisation Standards. These includes the standards for Markup Languages, Metadata Resources, Character Coding, Access Protocols and Interoperability, Content Creation, Manipulation & Maintenance, Authoring Standards, Text and Content Markup, Translation Standards, Terminology and Lexicography Standards, Terminology Interchange Standards, Controlled Language Standards, Taxonomy and Ontology Standards, Corpus Management Standards, Locale-Related Standards and Digital library standards.

Introduction:

Computer users working with multilingual text data face many challenges, especially when non-Roman scripts are involved. Commercial operating system and application software is evolving and becoming more sophisticated in handling multilingual data. To support users in different languages, programs must not only use translated text, but must also be adapted to local conventions. These conventions differ by language or region and include the formatting of numbers, dates, times, and currency values, as well as support for differences in measurement units, text sorting order, and other services.

Software Localisation is adopting a product to meet the language, cultural and other requirements of a specific target market especially in context of configuring an operating system or application programme with its character sets, date and time formats, currency formats etc.

Following discussion highlights the various Language Technology Standards and the organizations involved in the development of such standards.

Language Technology Standards

A. Markup languages:

A.1 Standard Generalized Markup Language (SGML):

An international standard (ISO 8879) published in 1986 that prescribes a standard format for embedding descriptive markup within a document, as well as specifying a standard method for describing the structure of a document.

A.2 Hypertext Markup Language (HTML):

A language for the exchange of scientific and other technical documents, suitable for use by non-document specialists, which addressed the problem of SGML complexity by specifying a small set of structural and semantic tags suitable for authoring relatively simple documents and providing support for hypertext, and which became the universal markup language used to create the World Wide Web.

A.2 Extensible Markup Language (XML):

A subset of SGML whose goal is to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with HTML and that has been designed for ease of implementation and for interoperability with both SGML and HTML.

Indic XML DBMS based on the level of semi-structured-ness of XML documents has been built with support for concurrency control. Current work is on implementing an Indic XML DBMS for handheld devices, such as, Simputers.

A.3 Extensible Hypertext Markup Language (XHTML):

An extension of HTML 4.01 that defines a single namespace for html. Extensible Hypertext Markup Language -A reformulation of HTML 4.0 in XML 1.0. XHTML is a new language for building web pages that has recently been proposed as a W3C Recommendation.

B. Metadata Resource

B.1 Resource Description Framework (RDF):

A language designed to support the Semantic Web, in much the same way that HTML is the language that helped initiate the original Web; a framework for supporting resource description, or metadata (data about data), for the Web. RDF provides common structures that can be used for interoperable XML data exchange. RDF follows the W3C design principles of interoperability, evolution, and decentralization.

B.2 Metadata Registry:

A family of standards that specifies basic aspects of data element composition, including metadata and that applies to the formulation of data element representations and meaning as shared among people and machines.

B.3 Unified Modeling Language (UML):

A graphic language for visualizing, constructing and documenting the artifacts of a software system.

C. Character Encoding Standards

C.1 ISCII:

The Indian Script Code Information Interchange was evolved by then Department of Electronics in 1988. This standard provides a common code and keyboard for 10 Indian scripts. The scripts are encoded according to a common goal: comparable characters are in the same order and the same relative location. This structural arrangement facilitates transliteration to some degree among Indian scripts. The Bureau of Indian Standards (BIS) adopted ISCII in 1991. ISCII is an 8-bit encoding that uses escape sequences to announce the particular Indic script represent by a coded character sequence.

C.2 UNICODE

An International Standard (fully synchronized with ISO/IEC 10646 in character repertoire and code points) that specifies a character coding system designed to support the worldwide interchange, processing, and display of texts written in the diverse languages of the modem world, as well as of texts written in classical and historical languages.

The current Unicode standard for Indian languages is based on the Indian Standard Code for Information Interchange (ISCII-1988). Thus, in any given Indic script, the consonant and vowel letter codes of Unicode are based on ISCII-1988.

Unicode supports ten major Indian languages and writing scripts: Hindi, Marathi, Gujarati, Tamil, Telugu, Kannada, Punjabi, Bengali, Oriya, and Malayalam. Currently fewer than a hundred languages are included in the Unicode.

D. Text and Content Markup:

D.1 XML Localization Interchange File Format (XLIFF):

A format whose purpose is to store localizable data and carry it from one step of the localization process to the other, while allowing interoperability between tools; supersedes Open Tag and differs in features and purpose from TMX.

E. Terminology and Lexicography Interchange

E.1 TermBase eXchange(TBX):

It is an ISO-16642 complaint open XML based standard format for exchanging terminological data while retaining complex embedding structures.

E.2 Terminology Markup Framework(TMX):

A meta model that defines underlying structures and mechanisms needed for the computer representation of terminological data that is designed to express interchange information between different TMLs (Terminology Markup Languages), but is not designed to serve as a TML itself.

E.3 Lexical Resource Markup Framework:

New work item proposal for ISO that defines underlying structures and mechanisms needed for computer representation of non-presentational lexical resource data designed to facilitate interoperability with terminological, NLP, and ontological resources.

E.4 Open Lexicon Interchange Format:

A format that focuses on the interchange of data among lexical base resources from various machine translation systems.

F. Corpus-related Standards:

F.1 Corpus Encoding Standard:

EAGLES/ISLE standard that specifies a minimal encoding level that corpora must achieve to be considered standardized in terms of descriptive representation (marking of structural and typographic information) as well as general architecture (so as to be maximally suited *for* use in a text database) and that also provides encoding specifications *for* linguistic annotation, together with a data architecture *for* linguistic corpora.

G. *Access Protocols and Interoperability*

G.1. Common Object Request Broker Architecture (CORBA):

An open, vendor-independent architecture and infrastructure that computer applications use to work together over networks

G.2 Simple Object Access Protocol (SOAP):

A lightweight and simple XML-based protocol that is designed to exchange structured and typed information on the Web and whose purpose is to enable rich and automated Web services based on a shared and open Web infrastructure.

H. Speech Standards

Standards for creating speech enabled applications, Text-To-Speech output, and communication protocols for speech on VoIP and Wide Area Networks are being evolved. Some of the speech standards being evolved are:

H.1 VoiceXML :

VoiceXML is a language for creating telephony-based, speech-user interfaces. It uses speech recognition and touchtone (DTMF keypad) for input, and pre-recorded audio and Text-To-Speech synthesis (TTS) for output. It is based on the Worldwide Web Consortium's (W3C's) Extensible Markup Language (XML), and leverages the web for application development and deployment.

H.2 SALT (Speech Application Language Tags): The Aurora Standard is a distributed speech recognition protocol defined within ETSI (European Telecommunications Systems Institute).

H.3 MRCP (Media Resource Control Protocol): MRCP (Media Resource Control Protocol) is offered as a standard communication protocol for speech resources across VoIP networks.

H.4 Aurora: The Aurora Standard is a distributed speech recognition protocol defined within ETSI (European Telecommunications Systems Institute). Aurora is focused on distributed speech recognition over the mobile cellular network.

H.5 SOAP (Simple Object Access Protocol): SOAP (Simple Object Access Protocol) is a lightweight protocol for exchange of information in a decentralized, distributed environment. It is an XML based protocol that consists of a framework for describing what is in a message and how to process it.

I. *Digital Library Standards:*

I.1 Dublin Core Standards:

The Dublin Core metadata element set is a standard for cross-domain information resource description. An information resource is defined to be "anything that has identity". There are no fundamental restrictions to the types of resources to which Dublin Core metadata can be assigned.

I.2 OCLC Standards:

Founded in 1967, OCLC Online Computer Library Center is a nonprofit, membership, computer library service and research organization dedicated to the public purposes of furthering access to the world's information and reducing information costs. More than 50,540 libraries in 84 countries and territories around the world use OCLC services to locate, acquire, catalog, lend and preserve library materials.

I.3 Information Retrieval Standard:

This standard was processed and approved for submittal to ANSI by the National Information Standards Organization.

Standards Development Organisations

The Free Standards Group is working to promote the use and acceptance of open source technologies through the application, development and promotion of interoperability standards.

In the International Standardization Organization is a network of national standards institutes from 140 countries working in partnership with international organizations, governments, industry, business and consumer representatives. ISO is working in development of so many language standards.

The Unicode Consortium is a non-profit organization founded to develop, extend and promote use of the Unicode Standard.

The World Wide Web Consortium (W3C) develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential as a forum for information, commerce, communication, and collective understanding.

OASIS is an international, not-for-profit consortium that designs and develops industry standard specifications for interoperability based on XML.

OLIF is XML-compliant and offers support for natural language processing (NLP) systems, such as machine translation, by providing coverage of a wide and detailed range of linguistic features.

The TermNet - the International Network for Terminology is promoting international co-operation in the field of terminology. The TMX and TBX standards are developed by OSCAR.

Conclusion

Government of India, Department of Information Technology (DIT) is the voting member of Unicode Consortium. Some discrepancies were observed in the Unicode for Indic scripts. DIT in consultation with Linguists, Indian IT Industry experts and respective state government proposed changes in the Unicode Standard for adequate representation of Indic scripts. Some of these proposed changes have already been incorporated in the Unicode Standard. Other issues related to Indic scripts are being taken up with concerned organization.

References:

http://www.unicode.org

http://www.tdil.mit.gov.in

http://www.lisa.org/standards/

http://appling.kent.edu/ResourcePages/LTStandards/Chart/LanguageTechnologyStandards.htm

Contact : tdiliufo@mit.gov.in