

ISO *Focus*

The Magazine of the International Organization for Standardization

Volume 4, No. 5, May 2007, ISSN 1729-8709

IT Quality and Security

- Infosys CEO Nandan Nilekani
- Watch your language! Terminology Standards



Developments and Initiatives

TC 37, Terminology and language

Commentators talk of the ubiquity of the Web, while discussions elsewhere concern the digital divide and the lack of content in a majority of the world's languages.

Although the predominance of use of the English language in readable Web content is gradually changing, a variety of studies demonstrate that the Web does not present a reliable surrogate for the use of languages in the world. This is possibly because the capability for representing these languages and dialects within these languages has been lacking.

Developments are currently beginning to emerge from ISO/TC 37, *Terminology and other language and content resources*, in respect of a significant expansion of the well-known series of International Standards, ISO 639, *Codes for the representation of names of languages*.

This is intended to provide a building block of basic identifiers (metadata) with which to index and retrieve the potential content of a truly diverse and multilingual information society. Previously, standards ISO 639-1 and ISO 639-2 provided for around 400 language identifiers.

Various estimates placed the number of languages in the world between 6 000 and 8 000, and the recently issued ISO 639-3 expands on the existing 400 to produce a set of over 7 500 language identifiers.

Languages are used in different ways, and some languages have a number of different ways in which they can be written, spoken or signed. Identification of these different ways pushes the total number of identifiers upwards of 30 000, and it becomes clear that until now the ISO 639 series has catered for a very small proportion of the true diversity of "languages".

The 639 series of ISO standards is generically titled *Codes for the representation of names of languages*, and is expected soon to be divided into parts similar to those shown in the table below.

At its most basic, the expansion from 400 metadata identifiers to upwards of 30 000 would provide support for a global thesaurus of names of languages, in every language of the world – some 400 million potential names, although the true figure is initially likely to be significantly lower.

The developments in the ISO 639 series offer the potential for the reuse of research materials which document the languages, the sum of which – the world's knowledge of the languages of the world – could be significantly greater than its parts.

ISO 639 forms a basic standard for many of the application areas for which ISO/TC 37 develops standards, and within which the ISO/TC 37 standards are used.

Work is already in progress in the Internet community through the Internet Engineering Task Force (IETF) to make use of these emerging standards, with a newer version of the IETF's language identification, incorporating ISO 639-3, expected shortly.

Historically, use of the IETF's output has been made by the eXtensible Markup Language (XML) – this development will eventually allow for new

Title of standard	Status	Registration authority	Number of identifiers (approx.)
ISO 639-1, Part 1: Alpha-2 code	Published (2002)	InfoTerm	150
ISO 639-2, Part 2: Alpha-3 code	Published (1998)	Library of Congress (LoC)	400
ISO 639-3, Part 3: Alpha-3 code for comprehensive coverage of languages	Published (2007)	Summer Institute of Linguistics (SIL)	7 000
ISO 639-4, Part 4: Implementation guidelines and general principles for language coding	Expected late 2007	n/a	n/a
ISO 639-5, Part 5: Alpha-3 code for language families and groups	Expected late 2007	TBC	100
ISO 639-6, Part 6: Alpha-4 representation for comprehensive coverage of language variation	Expected early 2008	GeoLang	25 000

Developments and Initiatives

contributors to make good quality identification within their XML documents.

There is interest, also, in the “Multilingual Internet” – described by some as a major element of the Next Generation Internet – being able to support domain names, e-mail addresses and other types of publicly readable protocol content in character sets other than ASCII.

The need for international country codes has been identified and a New Work Item Proposal submitted to ISO/TC 46/WG 2, *Coding of country names and related entities*. The project, proposed by BSI, aims to establish a joint working group between ISO/TC 37 and ISO/TC 46 and to set up liaisons with interested external organizations.

Further potential exists for these standards to support future generations of current Web-based technologies. For example, a future generation of search engines along the lines of Accoona, that already allows specialization of search by language identifier, might enable searches to be specialized for specific written forms of languages. Future versions of the video sharing website YouTube might allow for searches to

accurately index the languages spoken in the video clips, and other so-called Web 2.0 applications could similarly benefit.

Developments within the ISO 639 series were discussed at the Standards for Global Business Conference held in Vienna 14-15 November 2006, at which developers involved with the ISO 639 series discussed collaboration with the OmegaWiki project, a community-based website for the documentation of information about languages.

OmegaWiki will support the collection and collation of information about languages by the communities that use them.

This process will assist the verification and validation of language information by the newly-formed World Language Documentation Centre (WLDC), enabling the registration authority of ISO 639 part 6, GeoLang Ltd, to ensure the application of a full verification and validation methodology for the identifiers.

The collaboration has been agreed initially for ISO 639-6, the most ambitious of these standards yet, having to map to the existing parts of ISO 639,

About the authors



Dr. Lee Gillam is a Research Fellow in the Department of Computing at the University of Surrey and Director of GeoLang. Involvement with ISO/TC 37

comes through the British Standards Institution (BSI), efforts supported, in part, by the European Union's eContent programme of research under the Linguistic Infrastructure for Interoperable Resources and Systems (LIRICS). Research interests and publications encompass metadata and ontology, knowledge understanding, and high performance computing.



Debbie Garside is Managing Director of GeoLang Ltd and CEO of the World Language Documentation Centre. Ms. Garside is also a member of the Multilingual

Internet Names Consortium (MINC) Board and a member of its Secretariat as well as a Wikimedia Foundation Advisory Board Member. She is Convenor of ISO/TC 37/SC 2/WG 1/TG 2, Editor of ISO 639-6 and Chair of BSI mirror committee TS/1/-1. She is a member of the Country Code Names Supporting Organization and Government Advisory Committee for Internationalized Domain Names (ccNSO-GAC IDN) Joint Working Group. Ms. Garside's research interests encompass internationalization, morphology and human genetic linguistics.

ISO/TC 37 Annual meeting

The annual meeting of ISO/TC 37 will be held August 11-18 2007 in Provo, Utah. There will a one-day conference on August 13 on the “Pragmatic Applications of Standards” to look at the wide variety of applications of standards terminology to fields such as information technology, e-commerce and a wide variety of other fields.

For more information on the meeting, please contact: Sue Ellen Wright, sellenwright@gmail.com

take account of existing systems and support the interoperability with such systems.

The World Language Documentation Centre has been formed as an association of world experts and will act as a gatekeeper between language communities and the ISO 639-6 Registry, GeoLang Ltd.

The schedule for publication of ISO 639-6 should enable its availability during 2008, a year proposed as the UN International Year of Languages. ■

For further reading see:

Dr. Lee Gillam, Debbie Garside, and Chris Cox, (2007) “Developments in Language Codes Standards”. In Georg Rehm, Andreas Witt, Lothar Lemnitzer (eds.) *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen*. Proceedings der GLDV 2007, Tübingen: Gunter Narr Verlag. Tübingen University, April 11.13, 2007, pp. 281.289. <http://www.gldv.org/2007>.

Dr. Lee Gillam, Debbie Garside, and Chris Cox, (2006) “Information volumes and linguistic diversity: meeting the challenges for content management”. 3rd International Conference on Terminology, Standardization and Technology Transfer (TSTT), 25-26 August, Beijing, People's Republic of China.

Accoona: <http://www.accoona.com>

YouTube: <http://www.youtube.com>

OmegaWiki: <http://www.omegawiki.org/>

WLDC: <http://www.thewldc.org/>

GeoLang: <http://www.geolang.com/>