ISO TC37/SC2 N344
UTC-L2/05-242

ISO TC37/SC2
Terminographical and lexicographical working methods
Secretariat: Canada (SCC)

Doc. Type:	Liaison report
Title:	**TC37/SC2 Liaison Report to the Unicode Consortium**
Source:	Peter Constable (liaison to/from the Unicode Consortium)
Date:	September 20, 2005
Status:	For review by Unicode committees: UTC, CLDR

Earlier this year, a liaison relationship was established between the Unicode Consortium and ISO TC37/SC2. This is the first liaison report from TC37/SC2 to the Unicode Consortium. Status of specific projects of particular interest to the Unicode Consortium will be reviewed. Background on TC37 and on TC37/SC2 and its other projects will also be provided.

There are certain projects in other sub-committees of TC37 that may be of particular interest to members of the Unicode Consortium. These are reviewed at the end of this report. To be noted in particular are projects on word segmentation.

Highlights of particular interest to the Unicode Consortium

TC37/SC2 sees its ability to encourage the participation of stakeholders in the standards development process as being a major factor in its market relevance. Thus, TC37/SC2 welcomes the participation of the Unicode Consortium in projects of common interest

The TC37/SC2 standards of most direct interest to the Unicode Consortium are the ISO 639 series of standards for language identifier coding. The two existing standards in this series, ISO 639-1:2002 (alpha-2 code) and ISO 639-2:1998 (alpha-3 codes) are current, with the code tables being maintained by the ISO 639-RA Joint Advisory Committee (JAC). The JAC processes on-going requests for additions to the alpha-2 and alph-3 codes. From January 2004 to August 2005, seven new entries were added to the alpha-3 codes, and there are new requests in process. At present, there are 204 entries in the alpha-2 code and 474 entries in the alpha-3 codes.[1]

There are four TC37/SC2 projects in progress to develop additional parts to the ISO 639 series. The one of most immediate interest to the Unicode Consortium is ISO/DIS 639-3, which aims to provide an alpha-3 code that covers all known human languages. The initial code table for Part 3 is being formed by augmenting the existing entries in Part 3 with the inventory of modern languages found in *Ethnologue 15th edn.* along with catalogs of ancient and constructed languages maintained by The Linguist List.

ISO/FDIS 639-3 will be submitted for ballot soon, and approval is expected by the end of 2005.

The code table for ISO 639-3 will be published on the Web by the designated registration authority, SIL International. They have prepared a draft version of the ISO 639-3 Web site. SC2 has authorized them to make the draft site publicly available for feedback on the organization and presentation of information in

[1] Of the 474 entries in ISO 639-2, over fifty are for collections of languages rather than individual languages.

the language code database prior to the publication of ISO 639-3. The Unicode Consortium and its members are invited to review the draft site and to provide feedback on the organization and presentation of information in the language code database for Part 3.[2]

Background on TC37 and TC37/SC2

TC37 *Terminology and language and content resources* has the following scope:

> Standardization of principles, methods and applications relating to terminology and other language and content resources in the contexts of multilingual communication and cultural diversity.

TC37/SC2 *Terminographical and lexicographical working methods* (hereafter SC2) has the following scope:[3]

> Standardization of terminological and lexicographical working methods, procedures, coding systems, workflows, and cultural diversity management, as well as related certification schemes

Its mission is to provide practical advice concerning terminology work flows, translation-oriented terminography and lexicography, cultural diversity management, certification schemes and assessment methods for terminology and translation through the publication of standards and the use of the Internet in order to meet the needs of its client audiences.

TC37/SC2 has five working groups focusing in various areas as follows:

- WG1: Language coding.
  Scope: standardization of coding systems for the representation of names of languages and writing systems.
- WG2: Terminography
  Scope: standardization of the application of terminology and terminography work methods to various fields of use.
- WG3: Lexicography
  Scope: Standardization of the structuring of lexicographical data
- WG4: Source identification for language resources
  Scope: to prepare standards on the identification of sources of written, oral, audio-visual and electronic language resources (in the form of documents and any other forms) and on the metadata related to this identification
- WG5: Requirements and certification schemes for cultural diversity management

At the recent SC2 plenary meeting in Warsaw, SC2 took a resolution to form a sixth working group:

- WG6: Translation and Interpretation Services

There are eight current ISO standards developed by SC2. Beyond the ISO 639 series, the following, which may be of particular interest to the Unicode Consortium and its members:

- ISO 12199:2000, *Alphabetical ordering of multilingual terminological and lexicographical data represented in the Latin alphabet.*

[2] The ISO 639-3 site is located at http://www.sil.org/iso639-3. Please note that the code tables published at that site are tentative and subject to change until the date of publication.

[3] This title and scope for SC2 reflect recent changes brought about at the TC 37 and SC2 plenary meetings in Warsaw, August 22–26, 2005.

- ISO 12616:2002, *Translation-oriented terminography*.

Current TC37/SC2 projects

TC37/SC2 has current projects for development of a number of new standards related to terminology, lexicography and language coding.

There are four projects underway to develop new parts to the ISO 639 series of language coding standards:

- ISO/DIS 639-3, *Codes for the representation of names of languages – Part 3: Alpha 3 code for comprehensive coverage of languages.*
- ISO/CD 639-4, *Codes for the representation of names of languages – Part 4: Implementation guidelines and general principles for language coding*.
- ISO/DIS CD-5, *Codes for the representation of names of languages – Part 5: Alpha-3 code for language families and groups.*
- ISO/WD 639-6, *Codes for the representation of names of languages – Part 6: Alpha-4 code for the comprehensive coverage of language variation.*

Details on the ISO 639-3 project were provided above.

ISO 639-4 will establish the general framework for the ISO 639 language codes. In the long term, it is anticipated that much of the text that is now in each of the other parts of ISO 639 but common across parts will be centralized into this document. A CD ballot for this project was recently completed.

ISO 639-5 will extend the set of Alpha-3 identifiers for collections four in ISO 639-2. A CD was recently approved, and a DIS ballot is expected before the end of 2005.

The objective for ISO/WD 639-6 is to develop an Alpha-4 code for languages and language varieties using the inventory published in *Linguasphere*. This project is still in the preparatory stage, though a CD may be ready for ballot before the end of 2005.

Other TC37/SC2 projects that may be of particular interest to the Unicode Consortium and its members include the following:

- ISO/CD 10241-1, *Terminological entries in standards – Part 1: General requirements*.
- ISO/WD 22128, *Requirements for terminographical and lexicographical products*.
- ISO/WD 23185, *Assessment and benchmarking of terminology holdings: General concepts and principles*.

A new work item proposal for ISO 10241-2, *Terminological entries in standards – Part 2: Nationalization of international terminology standards* is in preparation.

TC37 projects outside of SC2

There are other projects managed by other sub-committees within TC37 that may be of interest to the Unicode Consortium and its members. Some of these are mentioned here briefly.

A current project of interest within TC37/SC3 is:

- ISO/CD 12620, *Terminology and other language resources -- Specification of data categories and management of a data category registry for language resources.*

This standard will apply the framework of the ISO/IEC 11179 family of standards to establish a registry of data categories for use across all TC 37 projects. This registry will be used to define reference categories for notions relevant to language resources thereby facilitating interoperability of resources created for different purposes.

TC37/SC4 was created with the objective to prepare standards that specify principles and methods for creating, coding, processing and managing language resources, such as written corpora, lexical corpora, speech corpora, dictionary compiling and classification schemes.

Several TC37/SC4 projects are designed to establish foundations for later projects:

- ISO/WD 24612, Linguistic Annotation Framework, will establish principles and mechanisms to be shared across the various data models developed within SC4.
- ISO/DIS 24610-1, *Feature Structures – Part 1: Feature Structure Representation*, will establish basic definitions and principles for XML representation of feature structures. Part 1 has recently been approved for publication as an ISO standard. A subsequent project is being initiated to develop Part 2: *Feature System Declaration*. The combined series will establish a reference data format for use in storage and exchange of feature structure information.
- ISO/WD 24611, *Morphosyntactic Annotation Framework*, will define elementary descriptors for morphosyntactic annotation needed in various types of linguistic models used for natural language technologies.
- ISO/AWI 24615, *Syntactic Annotation Framework*, will define elementary descriptors for syntactic annotation that will provides ways of representing syntactic information associated to a group of morpho-syntactically annotated items, independently of the processes that have led to the creation of this information.
- ISO/WD 24613, *Lexical Markup Framework*, will establish a general framework and meta-model for modeling lexical data as used in a wide variety of natural language technologies.

Two new SC4 projects have were approved earlier this year that focus on word segmentation, particularly in relation to languages for which word boundaries in text cannot be determined by character properties such as white space (e.g., Chinese, Thai):

- ISO/AWI 24614-1, *Word Segmentation – Part 1: Principles and Methods*
- ISO/AWI 24614-2, *Word Segmentation – Part 2: CJK*

These standards will aim to define segmentation as needed for a wide variety of applications such as proofing, information retrieval, machine-aided translation, text-to-speech pre-processing, etc. Part 1 will establish general principles and methodology. Part 2 will specify word segmentation for CJK. Additional parts are anticipated that focus on other languages, such as Thai.

The plan is that these standards will establish normative requirements for word segmentation to which software products should conform. None of the countries participating have indicated plans to make these requirements for software procurement or for marketing of software products within their countries. This possibility exists, however, and it should be assumed that the specifications will be formulated into national standards of the various countries involved.

Under the proposed schedule for ISO/AWI 24614-1/-2, a complete draft of Part 1 is expected by 2005-10-31. A draft for Part 2 ready for sub-committee review is not expected until September 2006. These drafts should be available on the TC37/SC4 Web site.[4]

[4] See http://www.tc37sc4.org.