ISO TC37/SC2
Terminography and Lexicography
Secretariat: Canada (SCC)

| | |
|---|---|
| Doc. Type: | Liaison report |
| **Title:** | **Liaison Report from the Unicode Consortium** |
| Source: | Peter Constable (liaison to/from the Unicode Consortium ) |
| Date: | August 22, 2005 |
| Status: | For review by SC 2 |

The Unicode Consortium[1] was formed for the purpose of developing an industry character coding standard that would be universal in coverage. Since 1991, the Unicode Consortium has worked in co-operation with ISO/IEC JTC1/SC2 on synchronized industry and ISO/IEC standards for a universal character set.

The Unicode Standard has been very successful, and has facilitated significant changes within information and communication technologies (ICTs) in relation to multilingual capabilities. Because many of the members of the Unicode Consortium are involved in the development of such technologies, it has been natural for the Consortium to take an interest in other work related to development of multilingual / multiregional / multicultural ICT products. This has led to additional standards or specifications developed by the Consortium, such as the Unicode Collation Algorithm and the Common Locale Data Repository Project, and to their support of related standards or specifications developed in the context of other industry bodies, such as tags for language identification.

The membership of the Consortium include organizations from several sectors, including software, localization and other ICT vendors; governmental bodies, universities, libraries and NGOs. The Consortium has been modifying the structure of membership categories in order to provide greater opportunities for organizations to be involved in the work of the Consortium. Participants in the work of TC37/SC2 are invited to review the new membership categories, details of which are available on the Consortium's Web site.

Recent developments in various areas of the work of the Unicode Consortium are of potential interest to TC37/SC2, and in some of these areas the Consortium would benefit from input from SC2 members or from experts with whom SC2 members may have contact. The following sections will provide a brief review of developments in the following areas: the Unicode Standard, the Unicode Collation Algorithm, the Common Locale Data Repository, and tags for language identification.

## The Unicode Standard

Version 4.1 of the Unicode Standard[2] was published earlier this year. In this version, the character set has expanded to include 97,720 characters covering 58 scripts and many kinds of symbols. The most notable changes are:

[1] Detailed information about the Unicode Consortium and its various standards, specifications and areas of work can be found online at http://www.unicode.org.
[2] Details regarding Unicode 4.1 can be found online at http://www.unicode.org/versions/Unicode4.1.0/.

- Addition of 1,273 new characters, including those needed to complete round-trip mapping of the HKSCS and GB 18030 standards, five new currency signs, some additional characters for Korean and scripts of India, as well as eight new scripts.
- New Unicode Standard Annexes — UAX #31, Identifier and Pattern Syntax, and UAX #34, Unicode Named Character Sequences — as well as significant changes to other Unicode Standard Annexes.

Much of the new work in the Unicode standard is associated with portions of the standard, or associated standards and technical reports, that specify the behavior of Unicode characters in processing, to facilitate ICT implementations and other ICT standards and protocols. For instance, a stability policy in relation to case mappings is planned for Unicode 5.0 in order to provide stability to standards and protocols for identifiers, such as Internationalized Domain Names. This has an impact on the way in which new characters are encoded: whenever uppercase characters are encoded, their lowercase equivalents should be encoded at the same time.

Other active work areas that supplement the Standard include:

- UTS #18, Unicode Regular Expressions — used for processing text
- UTS #22, Character Mapping Markup Language — used for describing hot to convert between different character sets
- UTR #30, Character Foldings (draft) — for transforming text in a variety of ways
- UTR #36, Unicode Security Considerations — for dealing with security issues, such as 'spoofing' and 'phishing'
- UTS #37, Registration of Ideographic Variation Sequences (proposed draft) —for registering different variants of CJK characters

Work is underway to prepare Version 5.0, which will add over 1300 new characters and add support for five additional scripts. Unicode 5.0 will be synchronized with Amendment 2 of ISO/IEC 10646:2003.

At this stage, most of the new characters being added to the Unicode Standard are for scripts used by particular language communities for whose languages there has been little or no support in ICTs up to now, or for historic scripts of primary interest to philologists and paleographers. For instance, scripts added to Unicode 4.1 include New Tai Lue (used in China), Tifinagh (used in northwest Africa) and Glagolitic; scripts that will be added in Unicode 5.0 include N'Ko (used in west Africa), Balinese (used in Indonesia) and Phags-pa (an historic script of China and Mongolia).

Characters also continue to be added for scripts that have already been supported in the Standard, including major scripts that have long been supported in ICTs, such as Latin and Cyrillic. For instance, ten new Cyrillic characters used for minority languages within Russia were recently accepted for encoding. Additional technical symbols, currency symbols and other characters are also anticipated.

Beyond character additions, other enhancements are also being to the Unicode Standard to facilitate ICT implementations and other ICT standards and protocols. For instance, a stability policy in relation to case mappings is planned for Unicode 5.0 in order to provide stability to standards and protocols for identifiers, such as Internationalized Domain Names.

## The Unicode Collation Algorithm (Unicode Technical Standard #10)

The Unicode Collation Algorithm, also designated Unicode Technical Standard (UTS) #10,[3] is a standard published by the Unicode Consortium that provides a specification for how to compare strings while remaining conformant to the Unicode Standard. It defines a process for collation that can be tailored to meet the requirements of different languages, and also a default collation table for all Unicode characters that products can choose to use.

[3] Details regarding UTS #10 can be found online at http://www.unicode.org/reports/tr10/.

In May of this year, a new version of UTS #10 was published. The most significant revisions included:

- enhancements to the specifications for language-appropriate searching and matching, including the introduction of notions of minimal versus medial versus maximal matching; and
- a small number of changes in the default collation table for Latin characters.

Future work items in relation to UTS #10 collation weights for all new Unicode characters, plus the establishment of a stability policy and processes for change management. The Consortium would particularly welcome feedback on these issues from TC37/SC2 and participants thereof.

## The Common Locale Data Repository Project

To support users in different languages, software systems must not only use translated text, but must also be adapted to local conventions. These conventions commonly include the formatting (and parsing) of numbers, dates, times, currency values, etc.; display names for language, script, region, currency, time-zones, etc.; collation order (used in sorting, searching, and matching text); identifying usage of measurement systems, weekend conventions, currencies, etc.; and so on. A set of such default conventions for a given user community (identified by language, script, region, variant, etc.) is commonly referred to as a *locale*.

The purpose of the Common Locale Data Repository project[4] is to provide a general XML format for the exchange of locale information for use in application and system development — the *Locale Data Markup Language* (LDML) —, and to gather, store, and make available a common set of locale data generated in that format — the *Common Locale Data Repository* (CLDR).

The Unicode Consortium recently released CLDR Version 1.3 and LDML Version 1.3. This release expands the repository to include data for a total of 296 locales, involving 96 languages and 130 regions.

Also introduced is a complete set of POSIX-format data generated from the XML format, along with tools for generating platform-specific versions.

The categories of data that can be stored were also expanded to include:

- UN M.49 region identifiers, including identifiers for continents and regions; and
- data to support localization of names for time zones.

The CLDR project would welcome input from language experts participating in the work of TC37/SC2 or with whom those participants may have contact regarding the CLDR supplemental data tables, which provide mappings between languages and scripts, and between languages and regions.[5] Both feedback on the data already compiled and new data for additional mappings would be welcome.

CLDR also provides information on the characters used by different languages. Input from language experts on this would be welcome as well.

## Tags for Language Identification

A significant number of ICT technologies and products make use of ISO 639 language identifiers indirectly by way of an IETF specification, RFC 3066: *Tags for the Identification of Languages*. This specification makes use of ISO 639-1 and ISO 639-2, but provides an extended syntax that allows for language varieties beyond those covered in ISO 639-1/-2, notably, regional dialects that can be distinguished using tags that combine language identifiers from ISO 639 with country identifiers from ISO 3166-1.

[4] Detailed information regarding the CLDR project is available online at http://www.unicode.org/cldr/.
[5] The supplemental mapping tables between language and script and between language and territory are available online at http://www.unicode.org/cldr/data/diff/supplemental/supplemental.html.

In recent years, it has been recognized that RFC 3066 needed to be extended to better cover ICT needs in relation to metadata elements that identify language varieties for various reasons. One of these is the need to distinguish between distinct orthographies in which a single language may be written (for instance, Chinese in simplified characters versus Chinese in traditional characters). Another is the need to constrain the syntax used for constructing tags so that processes can determine what type of information each element in a tag provides.

A project to make significant revisions to RFC 3066 is in its final stages. The Unicode Consortium has endorsed this project, and members of the Consortium's technical committees have been encouraged to participate in the IETF process in support of the project.[6] The Unicode CLDR project plans to make reference to the new specification once it is completed.

In accordance with IETF practice, the revised specification will have a new designation using a different number; it will not be called "RFC 3066".

The revised specification will continue to make reference to parts 1 and 2 of ISO 639. Plans have been made for a subsequent revision at a later date to incorporate ISO 639-3, which cannot be done until ISO 639-3 is published as an international standard. Completion of ISO 639-3 is viewed as urgently needed so that it can be incorporated into the IETF specification as soon as possible. The Unicode Consortium, therefore, wishes to encourage TC37/SC2 in advancing that project to successful completion as soon as possible.

The revised specification allows for the possibility of incorporating part 6 of ISO 639 at some point in the future. No specific plans or decision have yet been made, however.

## Review of particular items for consideration by TC37/SC2:

The following are those particular items mentioned above for consideration by TC37/SC2, its members, or individuals participating in its work:

- The Unicode Consortium invites participants in the work of TC37/SC2 or the organizations with which they are affiliated to familiarize themselves with the new categories of membership in the Unicode Consortium that are available.[7]
- Feedback on the Unicode Collation Algorithm standard is solicited, particularly in relation to future work on a stability policy and processes for change management.
- The CLDR project would welcome input from language experts participating in the work of TC37/SC2 or with whom those participants may have contact regarding the CLDR supplemental data tables that provide mappings between languages and scripts, and between languages and regions,[8] and also regarding information on the characters used for writing different languages.
- The Unicode Consortium wishes to encourage TC37/SC2 to advance ISO 639-3 to successful completion as soon as possible.

[6] IETF procedures allow for participation in IETF projects by individuals only, with each participant representing only themselves, not any company, organization or user community. Thus, neither the Unicode Consortium nor its organizational members can participate in an IETF project directly.
[7] Details on membership levels are available online at http://www.unicode.org/consortium/levels.html.
[8] The CLDR supplemental data tables are available online at http://www.unicode.org/cldr/data/diff/supplemental/supplemental.html.