Title:	**Comments on plan to restructure ISO 639**
Date:	February 17, 2010
Doc. Type:	Liaison Contribution
Source:	Unicode Consortium P.O. Box 391476 Mountain View, CA 94039-1476 USA http://www.unicode.org
Action:	For consideration by ISO TC 37/SC2, TC 37/SC 2/WG 1, ISO 639-RA/JAC; response requested from SC2
References:	ISO/TC 37/SC 2/WG 1/N181 ISO/TC 37/SC 2/WG 1/N184 ISO/TC 37/SC 2/N492 R-2 (= TC 37/N 622 R1) ISO/TC 37/SC 2/N508 (=TC 37/SC 2/WG1/N187) ISO/TC 37/SC 2/N513 ISO/TC 37/N607 (= TC 37/SC 2/WG1/N182) ISO/TC 37/N627
Distribution:	TC 37/SC2, TC 37 WG1, ISO 639-RA/JAC

This document provides comments from the Unicode Consortium on two topics related to maintenance of ISO 639: (1) plans to restructure ISO 639, and (2) maintenance procedures related to macrolanguages.

A. Plans to restructure ISO 639

On reviewing reports from the 2009 TC 37 meetings in Bogotá, it has come to the attention of the Unicode Consortium that TC 37/SC 2 is working with ISO CS to restructure the ISO 639 family of standards as a single standard with data published as a database.

> Following the endorsement of Annex ST to the ISO Supplement to Part 1 of the ISO/IEC Directives, allowing the publication of standards as databases, it has been decided that the ISO 639 standards series shall migrate to that publication format. This development will have a significant impact on the structure and maintenance of the ISO 639 series.
> (ISO/TC 37/N607, p. 3)

The ISO 639 standards are of vital importance to the Unicode Consortium and its members, and so we take great interest in any developments that impact the structure and maintenance of these standards.

We have not seen any documents describing what exactly is meant by publishing “as a database”. The code tables for ISO 639 have in the past been provided by the ISO 639-2 and ISO 639-3 registration authorities as machine-readable data files, which may be downloaded at no charge. We strongly request that this continue to be true. If the plan is to create a database with a richer, more informative data schema, this may be of interest. Note, though, that there may be software implementations that make use of the current file formats published by the RAs; thus, we would appreciate opportunity to review and comment on proposed specifications for revised data files several months in advance.

The meeting reports also indicate that the planned changes will have a significant impact on the maintenance of the ISO 639 language codes.

> We need to be clear and make people as well as stakeholders understand that language coding is a socio-linguistic matter that needs to be dealt with in an appropriate way. The Maintenance Team will have to be larger, and roles need to be well defined and distributed.
>
> A generic approach is needed to forward the process. We are also looking at the possibility of allocating a new Secretariat to validate 639. The possibility of this Secretariat being affiliated to an organization is also studied. Max Planck Institute has been approached and seems to be willing to assume this task.
> (ISO/TC 37/SC 2/N508)

The Unicode Consortium urges TC 37/SC 2 to keep in mind the broad use of ISO 639 in information systems and data as a new maintenance model is considered. We feel strongly that maintenance of ISO 639 should not be approached primarily with a methodology suited to academic research but rather with one suited to maintenance of widespread information systems.

We understand that the comprehensive nature of ISO 639 entails the coding of many not-well-documented languages, and that changes in the language codes are needed as such languages become better documented. Any changes to the language codes need to be handled with great care, however: ISO 639 language codes are widely used in Internet protocols and algorithms, in widely-deployed software systems, and in countless databases and documents. A single change, if not handled carefully, has the potential to cause widespread and costly harm. In creating a new maintenance team, there will be a challenge to instill an appropriate respect for these concerns. The larger the maintenance team, the greater that challenge will be.

In view of our concerns, the Unicode Consortium, as an A-Liaison to TC 37/SC 2, kindly requests that we be kept advised of developments in this work, including being given notice of relevant documents and meeting invitations. Notifications may be communicated to the Consortium via our liaison office, Peter Constable (petercon@microsoft.com), or via our office (details above).

In the latter regard, we note in document WG1/N184 that a meeting was held on March 31, 2009 to discuss these changes to ISO 639 structure and maintenance. The Unicode Consortium learned of this meeting only after reviewing meeting reports from the 2009 TC 37 meeting, many months after the meeting occurred. As a category A liaison, we would have expected to be notified of such meetings (see ISO/IEC Directives, Part 1, 1.18.2.1). We see that a meeting invitation and agenda was prepared (WG1/N180), but the Consortium was not notified of this.

We note the following, in document WG 1/N184:

> All formal decisions have been taken to move the entire ISO 639 series into the ISO Concept Database (CDB), to merge the code tables into one, and to maintain the code tables according to the procedures of Annex ST to the Directives (“Standards as databases”).

This statement is confusing: we have not been able to find even relevant NWIP documents describing planned changes to ISO 639, let alone relevant drafts of a new standard. Thus, it is unclear how it could

be that “all formal decisions have been taken”. Clarification of the status of this work is, therefore, requested from SC2.

We also note in SC 2/N513 the following (emphasis added):

> The rest of the meeting mainly consisted in brainstorming discussions regarding the implementation of the 639 Language coding series within the ISO Concept Database with a view to find a way to make the transition process as easy and transparent as possible to all parties involved. *A working document to that effect will be drafted by H. Hjulstad by the end of September 2009.*

We have not been able to locate this document in the TC 37, SC 2 or WG 1 document registers. We request notification when this or any other documents related to this planned work are available.

In summary, the Unicode Consortium has great interest and concern in any significant changes being considered that may impact the maintenance of ISO 639 and the structure of ISO 639 data files and requests to be kept closely informed and given ample opportunity to comment on and contribute to this plan of work.

B. Maintenance procedures related to macrolanguage scope

With the development of ISO 639-3, the notion of *macrolanguage entity* was introduced. The need to designate some previously-existing language entities as *macrolanguages* at the time that the comprehensive code set of ISO 639-3 was introduced is understood. The Unicode Consortium questions, however, the need to create *new* macrolanguage entities, and has concerns regarding changes to existing coded-language entities to change the scope *to macrolanguage* from some other scope.

In relation to change request 2009-048 processed by the ISO 639-3 RA, the entry “Latvian” (identifier = lv / lav) has changed from having *individual language* scope to that of *macrolanguage*. A consequence of this is that the identifiers lv / lav are now explicitly documented as encompassing the Latgalian language in addition to Standard Latvian. We are concerned that some implementations of ISO 639 may have assumed that lv / lav denote specifically Standard Latvian, and that for such implementations this change may have disruptive implications.

Note two aspects of the newly-established situation for Latvian varieties that may be problematic for implementers:

- There are alternate identifiers that can be used for Standard Latvian (lv / lav, lvs) or for Latgalian (lv / lav, ltg).
- The identifiers lv and lav, which have been in widespread use for many years, are now explicitly ambiguous.

This concern applies to any existing individual-language entity in ISO 639, particularly those that are in widespread use.

Thus, we request that TC 37/SC 2 and/or the parties involved in maintenance of ISO 639 code tables (currently, the RAs and the Joint Advisory Committee) consider freezing the set of macrolanguage entities to those that are currently defined.

Consider two situations in which a macrolanguage entity is likely to arise:

- Certain not-well-documented varieties become better understood with the result that what was once assumed to be a single language, with identifier xyz, is now understood to be two or more distinct, individual languages. As a result, the existing entry xyz is changed from *individual language* to *macrolanguage*.

- A developed language, with identifier xyz, has certain less-developed varieties associated with it. One of these less-developed varieties is found to be a distinct language and is coded. In addition, a new identifier is coded for the developed, "standard" variety, and the existing entity xyz is changed from *individual language* to *macrolanguage.*

In the former situation, since the varieties were so little understood, it is likely that there are very few records / documents for this variety and, hence, that there is very little if any usage of the existing identifier xyz. Thus, it is probably not problematic to simply create new entities for the distinct individual languages and to deprecate usage of the existing identifier xyz. (It is assumed, of course, that coded entities must never be completely removed or for identifiers to be re-assigned.) Continuing to recommend use of the existing identifier only invites problems of multiple representation and ambiguity, as described above.

In the latter situation, the existing identifier is likely to be in widespread use. For that reason, the change described to the existing identifier entails significant risk for implementations.

In both situations, therefore, we believe that change in scope of an existing entry to *macrolanguage* can and should be avoided.

As a corollary, we suggest that existing entries for developed languages should be explicitly understood as denoting the well-known and developed variety, and not less-developed varieties that may qualify as distinct languages. For example, de / deu / ger should be understood to denote Standard German and not Germanic varieties that are candidates for coding as distinct, individual languages. This should be explicitly documented, and implementers of ISO 639 should be given a recommendation to request addition of distinct languages (e.g., Latgalian) before incorporating such usage with the identifier for a related, more developed variety (e.g., Latvian).

Changes to the language code that involve introduction of a completely new entity with macrolanguage scope (i.e., no existing entity is changed in scope to become a macrolanguage) also impacts the entities that become encompassed by the new macrolanguage. (By definition, there must be some such entities.) If all of the entities involved are newly-created at the same time, this will not impact existing usage, though given the comprehensive nature of ISO 639-3 today it seems likely that any newly-created macrolanguage would have to impact some already-existing entities. Thus, even in these situations the introduction of a macrolanguage should be carefully considered, and it may be best to avoid this practice as well.