ISO / TC 37 / SC 2 / WG 1

TC 37 – Terminology (principles and coordination)
SC 2 – Layout of vocabularies
WG 1 – Coding systems

Convenor: **Håvard Hjulstad**

Document:	ISO / TC 37 / SC 2 / WG 1 **N 69**
ISO / TC 37 / SC 2 / WG 1 “Coding systems”	
Subject:	**Language coding**
Prepared by:	Håvard Hjulstad (convenor of ISO / TC 37 / SC 2 / WG 1)
Date:	2001-01-31

This document is a response to the decisions made during the TC37/SC2 meetings in London, August 2000. The convenor of SC2/WG1 was asked to look into the feasability to standardize language coding beyond the current coding in ISO 639-1 and ISO 639-2, and to make concrete proposals for work items to be carried out within TC37/SC2/WG1.

This document describes three (closely interlinked) projects. The third project only aim directly for International Standard.

1. Maintenance of a database of language coding, (extracts of) which should be freely available on the web.
2. Adding to the database languages that are currently not included in ISO 639-1 or ISO 639-2, without assigning standardized identifiers.
3. Development of an International Standard for coding mechanisms for language variation, including variation through time, geographically determined dialectal variation, writing system, etc.

Note on terminology: This document uses “ISO 639” not in the sense “ISO 639:1988”, but in the sense “ISO 639-1 + ISO 639-2”. For clarity “ISO 639-1 and ISO 639-2” is sometimes used.

Language coding database

The convenor of TC37/SC2/WG1 has already developed an internal database (using MS Access 97) which includes all items that are given standardized identifiers in ISO 639-1 and ISO 639-2, items that have been proposed for inclusion in the standard, and some (more or less random) additional items.

The database is available for downloading at http://www.rtt.org/ISO/TC37/SC2/WG1/639/ISO639.zip. (The file needs to be unzipped and opened with MS Access 97 or MS Access 2000.)

The tables for ISO 639-1 and ISO 639-2 may be generated from this database.

In addition, the database includes a lot of linguistic and other related information, which has been collected by the WG1 convenor from a variety of sources.

I would like to propose that this becomes the “official” database for the Joint Advisory Committee (JAC) to the Registration Authorities to ISO 639. The “ISO 639” tables are currently available from a large number of locations on the web. There is a great danger of errors being published. A common database as an “official source” is the only way to ensure consistency.

The current convenor of WG1 is volunteering to maintain this database in close co-operation with the JAC and the Registration Authorities. Members of WG1 will be informed of updates to the database, and they are invited to contribute.

Additional languages

A lot of user feedback indicates that many users need codes for a large number of languages that are missing in the current ISO 639 (even Part 2). Currently, the tables of ISO 639-1 contains 180 entries, and the tables of ISO 639-2 contains 438 entries. The database currently contains a total of 493 entries. (Count as of 2001-01-31.)

The additional languages will serve as a “waiting list” for possible future standardization, and it may have the effect of speeding up the standardization process. However, the normal procedures of addition to ISO 639 need to be followed (proposal, justification, JAC decision).

The convenor of WG1 is volunteering to update the database with additional entries and information as submitted. Additional information may also include language names in additional languages and additional sets of language identifiers (for reference).

Coding of language variation

Since this is the only project that intend to result in an International Standard (or possibly more International Standards), this point will present the only need for a formal New Work Item Proposal.

The present standard allows for language identification using the country identifiers and country subdivision identifiers in ISO 3166, Parts 1 and 2. For example: “en US” = “English in the USA”, “en CA” = “English in Canada”, “en US-CA” = “English in the state of California”. There is no standardized mechanism to specify, e.g. “English in North America” or “English in southern California”. The current standards give a highly inadequate tool to identify dialects.

There is under finalization an International Standard for scripts (ISO 15924), and it will be natural to combine the script identifiers with language identifiers, e.g. “ku Cyrl” = “Kurdish in Cyrillic script”, and “ku RU Cyrl” = Kurdish in Russia in Cyrillic script”. ISO 15924 is sufficiently detailed to meet the needs in this respect, and a registration mechanism has been set up to ensure that any shortcomings will be remedied.

There exists one identifier in ISO 639-2 with a somewhat special status: “sgn” = “Sign languages”. There is clearly a need to develop a deeper encoding structure for this item, as most or all of the sign languages have a clearly expressed relation to one particular spoken and written language, possibly with local variation which may depend on geographical and non-geographical factors. The development of the structure for this item should be viewed in conjunction with other developments of ISO 639.

There are a few language identifiers for diachronic stages of languages, e.g. “ang” = “Old English (ca. 450-1100)”, “grc” = “Ancient Greek (to 1453)”. However, there is currently no adequate general mechanism for identifying diachronic stages.

For many linguistic applications it would be highly desirable to have a formalism to express the hierarchy of language families. Although some details in the hierarchy are quite controversial among linguists, most of the hierarchy has general acceptance, and should be suitable for international standardization.

The current ISO 639-2 has a number of code elements for “other languages” within a linguistic group, e.g. “fiu” = “Finno-Ugrian (Other)”, meaning in effect all Finno-Ugrian languages that are not given separate identifiers: Estonian (et/est), Finnish (fi/fin), Hungarian (hu/hun), Komi (kv/kom), Mari (chm), Northern Sami (se/sme), Sami (smi), Votic (vot). In fact ISO 639-2 does not give this information explicitly, since no linguistic classification information is included in the standard itself. The item “smi” = “Sami” is to be interpreted as “Sami languages other than Northern Sami”.

Consequently, there are no real “group identifiers” in ISO 639, apart from those few groups that have no separate language identifier for any member of its group, e.g. “bai” = “Bamileke languages”.

There may be a need to identify, e.g. “Norwegian Nynorsk” as “nn/non” and “Norwegian” (“no/nor”) and “North Germanic” and “Germanic” and “Indo-European”. This may be achieved by establishing a special arrangement of the identifiers of ISO 639 in a hierarchical table, adding separate identifiers at those nodes in the network structure where identifiers are missing.

Conclusion?

This (the last section of this document) may call for two new parts of ISO 639:

ISO 639-3: A linguistically based hierarchical structure of the language entries that are included in ISO 639-1 and ISO 639-2 with, e.g. four-letter identifiers for nodes in the structure (i.e. language families and groups).

ISO 639-4: An International Standard specifying mechanisms for the identification of variants of languages, including geographical variants, temporal variants, and variants relating to writing system.

The convenor of TC37/SC2/WG1 plans to have first working drafts for both these items ready before the TC37 meeting week in August 2001, unless feedback to this document gives reason to proceed in a different direction.

Please submit feedback as soon as possible.