Document:	ISO / TC 37 / SC 2 / WG 1 **N 89**
ISO / TC 37 / SC 2 / WG 1 "Coding systems"	
Subject:	**Future development of ISO 639**
Prepared by:	Håvard Hjulstad
Date:	2002-03-04

Background, basis

ISO 639-1 (alpha-2 code)[1] and ISO 639-2 (alpha-3 code)[2] are designed to meet the needs of terminology and library applications. The two parts of the standard and the coordinated effort to develop these two parts represent a vast step toward a universally acceptable set of identifiers for linguistic units.

In particular the library community has a genuine need to keep the set of identifiers stable. There are at least a nine-digit number of records using these identifiers. Although there is broad acceptance that the present parts of ISO 639 will be developed further, this development needs to be conservative.

For the ICT industry and for language resource and language technology applications there is also a genuine need to expand the current set of language identifiers and language identification mechanisms greatly. There may be a need for identifiers for 15–20 times as many linguistic units as the current tables provide.

ISO / TC 37 is ready to initiate projects to meet these needs. The projects will be carried out within the framework of ISO / TC 37 / SC 2 / WG 1. It is, however, recognized that it may be necessary to utilize working procedures and organizational structures that are different from most projects under ISO / TC 37 and other ISO committees. It will not be possible to meet the requirements as to timeliness without substantial external funding.

Outline of projects

The following projects are closely interconnected. It will be necessary to prioritize. For most projects deliverables may be subdivided so as to accommodate specified needs as rapidly as possible.

1. **A model for language identification**, including operational definitions of "language", "individual language", "language variant", "dialect", etc. The model will form the theoretical basis for future development. However, it is not deemed necessary that the model description be finalized in all detail before any other activity is initiated. The final deliverable will probably not be an International Standard. If there is a need for publication as an ISO document, it may be feasible to draft a Technical Report. Some of the results of this project may be incorporated in the deliverable from the following project.
2. **Language identification structure**. It is anticipated that language identifiers in the form of alpha-2, alpha-3, and alpha-4-5 code elements will not be the only mechanism needed for language identification. There will be a need to specify some or all of the following types of information in a number of combinations: geographical variation, variation as to script, writing system, and orthography, temporal variation, stylistic variation, etc. Structures for these language identification mechanisms need to be specified, possibly in varying formats depending on applications. The deliverable from this project may be an International Standard (possibly a new part of ISO 639).

---

[1] ISO 639:1988, *Code for the representation of names of languages,* and its next edition, currently ISO/FDIS 639-1 (2002), *Codes for the representation of names of languages – Part 1: Alpha-2 code.*

[2] ISO 639-2:1998, *Codes for the representation of names of languages – Part 2: Alpha-3 code.*

Håvard Hjulstad
Solfallsveien 31 tel: +47-64944233 mob: +47-90145563
NO-1430 Ås, Norway email: havard@hjulstad.com

3. **Linguistic unit description format**. It will be crucial to record linguistic and non-linguistic information about each language. To enhance the usability of this information within the project, and also to facilitate various types of external use for the information, a suitable structure needs to be developed. The deliverable from this project will not be published as an International Standard.

4. **Description of linguistic units and default values**. Based on the description format all relevant information needs to be recorded for all items that currently have identifiers assigned, as well as for all new items. This includes attributes such as script, orthography, and geographical area of use. To reduce the need for applications to specify all values, a set of default values need to be given as a part of the publication in electronic form. The documentation needs to be made available for all users in electronic form, but it is not anticipated that this information will be printed in the International Standards. The format of the specifications and the default value information needs to be established in close cooperation with developers of applications. It needs to be emphasized that this is a very large project, which needs to be subdivided into sub-projects with specific targets. It also needs to be emphasized that the quality of this work will be essential for the overall quality assurance for all related projects.

5. **Resolution of problems in current code tables**. Studies made by experts at SIL International[3] and other bodies have revealed errors and potential problems within the current ISO 639-1 and ISO 639-2 tables. The problematic items need to be clearly identified and brought to the attention of the Joint Advisory Committee (JAC)[4] to the ISO 639 Registration Authorities[5]. The JAC needs to process each individual item as a change proposal. All internal problems need to be resolved without unnecessary delay.

6. **Further development of ISO 639-1 and ISO 639-2**. Proposals for additional language identifiers in the alpha-2 and alpha-3 code lists will be processed by the JAC as received. It may be desirable to submit proposals to the alpha-2 and alpha-3 code lists in conjunction with the project that is described in this document. (Items in the alpha-4 code list described below may be candidates for later inclusion in the alpha-2 and alpha-3 list.)

7. **Hierarchical language identifiers**. The current ISO 639-2 includes some "language group identifiers". However, some users need a number of additional identifiers. This must be based on a pragmatic approach. The number of new identifiers will probably be a high two-digit or a low three-digit number. It should be investigated whether it is feasible to process these identifiers through the JAC and assign alpha-3 identifiers. There are at least two alternatives to this: One alternative is to assign alpha-5 identifiers to these items. In that case the alpha-5 code list should probably be published as an International Standard as a separate part of ISO 639. A second alternative is to incorporate these items in the alpha-4 code list described below.

8. **Additional individual language identifiers**. An estimate of 5000–7000 additional languages need to be registered. Each individual language will be assigned an alpha-4 identifier. Material that has already been collected and published by SIL International (the *Ethnologue* database[6]) and the Linguasphere Register[7] will be utilized as appropriate following the model and the structure of this project. The deliverable will be an International Standard as a separate part of ISO 639. If it is deemed useful, preliminary versions may be published as Technical Reports. The project needs to record in a systematic way linguistic and non-linguistic information that will be better suited for electronic publication. This should be done in the form of a publicly available database. (See also item 4.)

9. **Geographical coordinate information**. Including geographical coordinate information as an alternative to or in addition to other types of geographical specification, will greatly enhance the possibility to use language information in conjunction with a topic map as described below. It will be desirable to include several layers of such information, e.g., core geographical area of use, area of majority oral use, area of minority oral use, area of official written use, area of use as a second

---

[3] http://www.sil.org/

[4] http://lcweb.loc.gov/standards/iso639-2/iso639jac.html

[5] ISO 639-1 Registration Authority: Infoterm, Vienna. ISO 639-2 Registration Authority: Library of Congress, Washington DC.

[6] *Ethnologue: Languages of the World,* http://www.ethnologue.com/

[7] The Linguasphere Register of the World's Languages and Speech Communities, http://www.linguasphere.org/

language, area of first language elementary school teaching, area of second language elementary school teaching, etc., bearing in mind that all such areas may be discontinuous. Deliverables from this project will not be published separately, but be included in other publications and presentations.

10. **Topic mapping project**. This will be a separate project utilizing expertise and technical resources other than those needed for the other projects. Various layers of electronic linguistic maps in combination with electronic maps including political and administrative boundaries, demographic information, and place names will form an extremely valuable tool for effective presentation of information that is collected and recorded in other projects.
11. **Mapping with other language identification code sets**. Extensive mapping tables need to be developed and published to document the relation to e.g. *Ethnologue* and the Linguasphere Register. This must be done in close cooperation with developers and maintenance authorities for other sets of language identifiers. These tables should be included in electronic publications, but may also be published in paper form, possibly as Technical Reports.

Organizational structure

Traditionally ISO / TC 37 projects have been managed by a Project Editor (often a Working Group convener) in continuous interaction with a Working Group. In the majority of cases Project Editors have only been able to devote a very limited part of their time to the project. For reasons that are easy to appreciate, contributions from other Working Group members are frequently very limited indeed during the time that the Working Group is not convened in meetings.

To ensure a timely operation of the language identification projects that are outlined in this document, it will be necessary to allow a Project Team consisting of, e.g., two or three experts to devote a major part of its time to the management of the projects. It is probably advisable that the Project Team report to both ISO / TC 37 / SC 2 and to a separate Steering Group.

The Project Team will need to cooperate closely with a larger Expert Group. The members of this group should also be members of ISO / TC 37 / SC 2 / WG 1, which will be the formal channel to the ISO system for the publication of ISO documents.

It is probable that there will be a need to establish a Registration Authority to keep the information up to date once the project is completed.

Cooperation with other activities

It is recognized that SIL International and the Linguasphere Register have already produced immensely valuable material that will provide direct input to the projects that are outlined in this document. Close cooperation with both these organizations will clearly be needed.

An initial analysis indicates that SIL International's *Ethnologue* has a theoretical basis and approach that corresponds very closely with the projects that are outlined here. Provided that a proper infrastructure is established, a merger of *Ethnologue* and the extended ISO 639 may be feasible, although it is much too early to assume that this will happen, let alone when and how this will happen. If possible, SIL International should be represented in the Project Team. It may also prove to be a practical solution to locate the anticipated new Registration Authority at SIL International, although it is much too early to make any decision in this respect.

The Linguasphere Register has a somewhat different approach. It may be better for both projects that the Linguasphere Register be allowed to develop independently of the extended ISO 639. However, Linguasphere Register expertise should be utilized, and the Linguasphere Register should be represented in the Expert Group that is mentioned above. Linguasphere experts have indicated a desire to retain reference to ISO 639 identifiers within the framework of their work.

Funding needed

Much more work is required to estimate the needed size and sources of funding.

It must be emphasized that the overall project and even many of the individual sub-projects may be subdivided in a number of ways depending on the urgency of specific needs and the availability of human, technical, and economic resources.

In addition to remuneration for the Project Team, funding will be needed to pay for external expertise, technical equipment and assistance, travel, as well as infrastructure costs.