ISO / TC 37 / SC 2 / WG 1

TC 37 – Terminology and other language resources
SC 2 – Layout of vocabularies
WG 1 – Coding systems

Convener: **Håvard Hjulstad**

Document:	ISO / TC 37 / SC 2 / WG 1 **N 72**
ISO / TC 37 / SC 2 / WG 1 “Coding systems”	
Subject:	**“Additional language coding” – A pre-WD**
Prepared by:	Håvard Hjulstad (convener of ISO / TC 37 / SC 2 / WG 1)
Date:	2001-07-18

Background

ISO 639-1 and ISO 639-2 include one mechanism to identify “language variety” by combining language identifiers with identifiers from ISO 3166 (all parts). However, this mechanism is highly inadequate. The standards do not specify clearly how the identifiers should be combined. The following examples have been seen:

en **term** /US/
en **term** US
enUS **term**
en US **term**
en-US **term**

Language variation exists on many more levels than geography. This includes temporal variation, sociolinguistic variation, and stylistic variation.

What can/should be standardized?

First negatively: Designations for specific dialects should not be standardized.

However, mechanisms for specifying linguistic variation may be suitable for standardization. The mechanism should be split into two aspects: internal representation and suggested presentation forms. The latter could be normative up to a point, e.g. for use in standardized vocabularies. Other usages may require different presentation forms, and the standard should allow this.

The following uses an “SGML-based” notation. The actual notation in the final document needs to be aligned with relevant SGML and XML applications.

This document uses the term “language tag” denoting a language identifier plus one or more attributes and attribute values.

Some details for a New Item Proposal

At least the following attributes may be defined (with random designations here): **geog** (geographical specification), **script** (writing system), **temp** (temporal specification), **socli** (sociolinguistic specification), and **style** (stylistic specification).

geog

For countries and country subdivisions the identifiers in ISO 3166 should be used. However, there should be a mechanism to describe larger and smaller areas. Standardization of such area identifiers should be left to possible new developments within ISO 3166.

Examples: geog=“CA+US” (Canada and USA), geog=“CA+US not US-HI” (Canada and USA not including Hawaii).

script

The script attribute should use ISO 15924.

Example: script="Latn" (Latin).

temp

Identification of time should use the common calendar.

Examples: temp="196X" (period from 1960 to 1969), temp="15XX" (the sixteenth century), temp="08XX-1255" (period from the ninth century to 1255), temp="b6XXX" (the seventh millennium BC).

socli

Sociolinguistic variation may be described in many different ways. The values of the **socli** attribute should probably be taken from an open list.

style

Stylistic variation should also have attribute values from an open list.

Defaults

It would be most useful to have a clear description of default values of each of the attributes for all languages that are included in ISO 639-1 and ISO 639-2, whenever such values exist. There will most likely not be any consensus of the default value of the **geog** attribute of an item like **en**/**eng**, but I should think that there is consensus of the default value of the **script** attribute of that item. That way it would not be necessary to specify the script for an English text, unless it was written in a script other than Latin. It will, however, be necessary to specify **geog** unless the text (term, word) is "unmarked" as to localization.

For the purpose of any application it would most likely be useful to specify defaults that are not universally true. Any dictionary may, e.g. state that terms with no **geog** attribute are valid for one particular country or for all countries or areas in a list.

Internal representation

The following notation is just a random "invention". It needs to be aligned with relevant notations.

<lang id="en" geog="AU" temp="18XX"> = Australian English of the 19th century.

<lang id="mis" geog="AU"> = Miscellaneous languages in Australia.

External presentation

I am uncertain as to how far it is useful to go when it comes to standardizing the presentation form. In ISO terminology standards the *Directives* specify the following: "en **term** US" (normally without the language identifier, since language is in most cases implicit from the layout).