SC22/WG20 N759

2000-05-22

From: John Clews

ISO 639: Language codes: report to ISO/IEC JTC1/SC22/WG20, 22 May 2000
[updated, 22 May 2000, from report to CEN/TC304 in (TC304.2292)

I represented CEN/TC304 at the ISO 639 Joint Advisory Committee in Washington in February 2000, which brings together experts from ISO/TC37/SC2 and ISO/TC46/SC4, who have developed the two parts of ISO 639 (Language codes). This is a report of that meeting, and on related issues that have surfaced subsequently. It also includes comments by Martin Duerst (W3C) and Erkki Kolehmainen (CEN/TC304) on some related issues.

The basic problem can be stated thus. IT systems have a need for specifying language codes. This summarises the current situation:

(a) 2-letter codes (ISO) - not enough code space available. Widely used.
(b) 3-letter codes (ISO) - enough code space available;
Widely used, but only in libraries.
not enough codes allocated;
slow to make new codes: long delays.
Also many national and other code variants.
(c) 3-letter codes (SIL) - enough codes allocated; no delays.
Quite widely used outside libraries.
SIL is in discussion with various
IT industry companies and with IETF.
Not standardised with ISO;
also some different entities involved
(spoken languages, dialects, etc.)

(d) RFC 1766 (IETF/W3C) - currently uses (a), not (b);
Succesor RFC to RFC 1766 currently
has considered (b) as well as (a),
but has not considered (c).
RFCs better known than ISO standards?

SIL has stated that it is willing to look at overlaps/conflicts, and to consider compromise. ISO has not discussed SIL at all.
One issue seems clear from this: SIL codes are currently more stable than ISO 639 codes and their variants. Appendix 1 notes JTC1/SC22/WG20's requirements for code stability. Appendices 2-5 provide information on some of the variants likely to get confused in specifications related to ISO 639-2.

Nobody has yet compared ISO and SIL lists to assess overlap or conflict, or checked on the feasiblity of combining them. John Clews plans to do this.

Providing input from ISO/IEC JTC1/SC22/WG20 into a replacement for RFC 1766 may also be of use.

This report covers the following areas:

1. Overview of the ISO 639 (language codes) standards

- 3-letter codes from ISO 639-2 (Language codes)
- 2-letter codes from ISO 3166 (Country codes)
- 4-letter codes from the draft ISO 15924 (Script codes)
- RFC 1766 (Language Tags) and its replacement

2. Meeting overview
- Harmonisation between ISO 639-1 and ISO 639-2;
- problems of variant codes;
- problems of coping with demand for new languages.

3. Notes on specific issues

4. Decisions on specific languages

Appendices which follow also note ISO/IEC JTC1/SC22/WG20's requirements for stability, and likely departures from this in national and other variants of ISO 639-2.

1. Overview of the ISO 639 (language codes) standards

ISO 639 (2-letter language codes) has existed for many years.
ISO 639-2 (3-letter language codes) is new, but its precursor (the USMARC language codes, maintained by the Library of Congress) has also existed for many years.
ISO WD 639-1 (revision of 2-letter language codes) is being developed and will replace ISO 639 if successful.

Official versions of these standards exist at the following web sites (thanks to Martin Duerst for this list):

The ISO 639-2 standard is at the official ISO 639-2/RA Web site:
http://www.loc.gov/standards/iso639-2/langhome.html

The tables arranged by ISO 639-2 code include the ISO 639-1 code as well:
http://www.loc.gov/standards/iso639-2/termcodes.html
(arranged by ISO 639-2/T (terminology) code)
http://www.loc.gov/standards/iso639-2/bibcodes.html
(arranged by ISO 639-2/B (bibliographic) code)

For further information see the ISO 639-2/RA home page:
http://www.loc.gov/standards/iso639-2/

2. Meeting overview

The ISO 639 Joint Advisory Committee meeting was successful in its own internal aims of ensuring compatibility between ISO 639 (soon to be replaced by ISO 639-1) and ISO 639-2, and of speeding up procedures for registration.

It also considered including all information in a single new ISO 639 standard, and discussed ways of relating the Internet standard RFC 1766 (and its proposed revision) to the ISO 639 language code standards, to the ISO 3166 country code standards, and to the draft ISO 15924 script code standards.

2.1 Relationship between ISO 639 standards and RFC 1766

The relationship between ISO 639 standards and the Internet usage involves enabling new code combinations in IT systems based on the

update to RFC 1766, by linking:
- 3-letter codes from ISO 639-2 (Language codes)
- 2-letter codes from ISO 3166 (Country codes)
- 4-letter codes from the draft ISO 15924 (Script codes)

Martin Duerst notes that:

> the successor of RFC 1766, currently in the works, will allow the use of 639-2 codes (3 letters) (for languages where not already 639 codes (2 letters) are in use. The successor of RFC 1766 will also, in the same way as 1766, allow additional registrations under the i- prefix, so that needs of the Internet community not covered by the various 639 codes can be dealt with.

Erkki Kolehmainen <erkki.kolehmainen@tieke.fi> on 3 May 2000 noted to CEN/TC304 in (TC304.2291) "Update on Language Codes" that:

> A list of recent changes to language identifiers defined in ISO 639 was published by the ISO 693-2 Registration Authority (the US Library of Congress) at http://lcweb.loc.gov/standards/iso639-2/codechanges.html
>
> The list includes a number of newly assigned two letter codes for some of the less common languages, including additional language variants for some European languages (Norwegian, Sami, Manx, etc.).
>
> A recently published IETF draft on Tags for the Identification of Languages at http://www.ietf.org/internet-drafts/draft-alvestrand-lang-tags-v2-01.txt will, when approved, allow both two and three letter ISO 639 language codes to be used as language identifiers within HTML and XML documents. The draft will obsolete RFC 1766, the current web reference document relating to the use of two letter language codes.

Martin Duerst (W3C) <duerst@w3.org> noted on 8 May in (TC304.2294) "Update on Language Codes" that

> The update of RFC 1766 will not immediately allow the use of the new codes in HTML and XML, because these reference RFC 1766. However, our intent at W3C is to make sure already now that it is clear that there should be no restriction e.g. in XML to hinder an adoption of the successor of RFC 1766, and we also will try to update these specifications (also CSS and others) as soon as possible.
>
> Where two-letter-codes are established, no 3-letter codes should be used in any Internet or Web document, to avoid confusion. This is part of the details of the RFC 1766 update.

However, in two other areas the ISO 639 Joint Advisory Committee was less successful:

2.2 variant codes:

Some very large users (such as Unesco, and the Linguist List web site) use 3-letter language codes from the Ethnologue, maintained by the Summer Institute of Linguistics (SIL). The 3-letter structure in both leaves room for ambiguity if the source of any 3-letter code is not described.

The ISO 639 Joint Advisory Committee failed to consider any ways of dealing with this issue.

In addition, there are various national variants of the ISO 639-2 codes, e.g. in the United States, the United Kingdom, Germany and Sweden: again the ISO 639 Joint Advisory Committee did not consider how this issue will be dealt with, and apparently there is no intention to describe any variant use in conjunction with ISO 639-2.

2.3 coping with demand

Given the large number of languages for which codes may well be required, and the length of time it took for the ISO 639 Joint Advisory Committee even to approve a small number of new language codes, and the 50-document restriction imposed by ISO 639-2 before requests for new codes will be considered, there are doubts on whether the ISO 639 Joint Advisory Committee will be able to keep up with the demand for codes.

By way of example, there remain many languages (some with official status in Europe) for which codes have been requested by CEN/TC304 for which no codes have been allocated, and decisions have been deferred, and no target dates have been specified for their reconsideration: requests for these were made over three years ago.

3. Notes on specific issues

3.1. ISO 639 (2-letter codes) may be slightly enlarged (i.e through adding more codes) as ISO 639-1, currently under ballot.

3.2. ISO 639-2 (3-letter codes) will be where most development will take place, and a larger number of codes can be expected to be allocated there. More codes will be allocated in ISO 639-2 than in ISO 639-1.

3.3. Unfortunately, CEN/TC304's request for codes was ignored, in effect, despite my protests, by considering the list, one by one, but in most cases deferring them to a future meeting (date unspecified).

There is also a requirement that codes would only be allocated in ISO 639-2 if there was evidence that there was 50 publications in that language. That might rule out some of those that CEN/TC304 requested, despite there being a need for such codes, e.g. on new websites for particular linguistic/ethnic groups, which are now springing up in Europe. Library of Congress representatives in particular were implacable in considering any change to that provision.

3.4. There may be other ways to provide for CEN/TC304's needs for codes for additional European languages: on the Internet, various users plan to use 3-letter SIL codes irrespective of what happens to ISO 639-1, ISO 639-2, or RFC 1766 (see 6 and 7 below).

SIL 3-letter language codes are used in various Unesco publications, and will also be used on the largest linguistic website, the Linguist List website. SIL has also been involved in discussions with participants in IETF, and with some vendors, about using information, and possibly codes, from SIL.

3.5. On the plus side, there was agreement that ISO WD 639-1 and ISO 639-2 will be developed in tandem, by the two different ISO committees, although there was talk of uniting them in a single standard (which I do not expect to come to fruition). Ensuring compatibility between the two standards, and the two committees was the main issue dealt with, rather than addressing user needs.

3.6. There was also discussion on how 2-letter and 3-letter codes from ISO 639-1 and ISO 639-2 wouould be used in RFC 1766, or rather its successor RFC. CEN/TC304 should aim to have input into that process, to ensure that whatever is produced as a result, the successor RFC provides codes for the language entities that CEN/TC304 has requested.

3.7. There was absolutely no discussion of 3-letter SIL codes: the convenor absolutely forebade the attendance of a SIL representative at the ISO 639 Joint Advisory Committee meeting, despite his presence in Washington.

3.8. The ISO 639 Joint Advisory Committee, with input from 3 members designated by ISO/TC37/SC2, and 3 members designated by ISO/TC46/SC4, plus two observers (Michael Everson and myself) seems to be the main developer of any future language codes, rather than any WG of ISO/TC37/SC2 or ISO/TC46/SC4.

3.9. ISO/TC46, the parent committee, is losing its secretariat, and by 8 May 2000, no country had offered to take it over. Some maintenance mode, overseen by ISO CS, is under consideration. How this will affect ISO/TC46/SC4, and the continued development of ISO 639-2 is unclear.

In passing, there also remain unresolved questions over who will provide the maintenance agency for ISO 3166: country codes, as the current maintenance agency gave this up., apparently earlier in 1999.

3.10 CEN/TC304, and also ISO/IEC JTC1/SC22/WG20, should also consider having discussions with IETF and SIL (the Summer Institute of Linguistics, based in Texas, but international in scope, with SIL offices in some European countries), about other ways to ensure that codes are available.

There are some incompatibilities between SIL codes and ISO 639-2 codes, but SIL does provide codes for most of the languages requested by CEN/TC304. SIL have also said that they are willing to amend SIL codes if this will assist in wider use.

3.11 ISO/IEC JTC1/SC22/WG20 also has needs for language codes and I will discuss this further there in May. Language codes may also come up in discussion at the next meeting of ISO/TC37/SC2 in London in August 2000. I am on the UK committee monitoring ISO/TC37.

JTC1/SC22/WG20 and CEN/TC304 should aim at a joint approach as there are similar needs.

I intend to do further work on this, and to contact the parties concerned, and to report back to CEN/TC304 (and to JTC1/SC22/WG20) by their next meetings.

4. Decisions on specific languages

This section lists the results of CEN/TC304's request for language codes, in terms of those accepted, rejected, or defered. In simple terms there are
(a) less European languages accepted than other languages (and then only into ISO 639-1, although Norwegian Bokmal and Norwegian Nynorsk were accepted into ISO 639-2);
(b) more European languages than other languages rejected; and
(c) more European languages than other languages defered.

4.1. Accepted (mainly alligning ISO 639-1 and ISO 639-2: few new codes were added. Only a few languages here were European).

4.1.1 European languages

> $ Bosnian: Accept into ISO 639-1; the code: "bs"/ "bos" for ISO 639-2
> $ Church Slavonic; Church Slavic; Old Slavonic; Old Church Slavonic:
   Accept into ISO 639-1; the code: "cu"
> $ Chuvash: Accept into ISO 639-1; the code: "cv"
> $ Komi: Accept into ISO 639-1; the code: "kv"
> $ Northern Sami; Sami, Northern: Accept into ISO 639-1; the code: "se"
> $ Norwegian Bokmal: Accept into ISO 639-1; the code: "nb" /
   accept into ISO 639-2; the code: "nob"
> $ Norwegian Nynorsk: Accept into ISO 639-1; the code: "nn" /
   accept into ISO 639-2; the code: "nno"

4.1.2 Other languages

> $ Avestan: Accept into ISO 639-1; the code is being reconsidered
> $ Chamorro: Accept into ISO 639-1; the code: "ch"
> $ Hiri Motu: Accept into ISO 639-1; the code: "ho"
> $ Kikuyu; Gikuyu: Accept into ISO 639-1; the code: "ki"
> $ Marshall; Marshallese: Accept into ISO 639-1; the code: "mh"
> $ Navajo; Navaho: Accept into ISO 639-1; the code: "nv"
> $ Nyanja; Chechewa: Accept into ISO 639-1; the code: "ny"
> $ Ossetian: Accept into ISO 639-1; the code: "os"
> $ Pali: Accept into ISO 639-1; the code: "pi"
> $ Sardinian: Accept into ISO 639-1; the code: "sc"
> $ Tahitian: Accept into ISO 639-1; the code: "ty"

4.2. Rejected

4.2.1 European languages

> $ East Frisian; Frisian, East; Sater Frisian; Frisian, Sater: reject
> $ Friulian: reject
> $ Frisian, North; North Frisian: reject
> $ Istro-Romanian: reject
> $ Kashubian: defer for ISO 639-2 / Reject for ISO 639-1
> $ Ladin: reject
> $ Ladino: reject

> $ Lallans; Lowlands Scots: reject
> $ Livonian: defer for ISO 639-2 / Reject for ISO 639-1
> $ Lower Sorbian; Sorbian, Lower: reject
> $ Mingrelian: reject
> $ Romany; Romani: reject
> $ Ruthenian; Rusyn: defer for ISO 639-2 / reject for ISO 639-1
> $ Upper Sorbian, Sorbian, Upper: reject
> $ Veps: reject

4.2.2 Other languages

> $ Efik: reject
> $ Mandingo: reject
> $ Old Persian; Persian, Old: reject

4.3. Deferred

4.3.1 European languages

> $ Abaza: defer
> $ Adyge: defer
> $ Aragonese: defer
> $ Aromanian; Arumanian: defer
> $ Arvanite: defer
> $ Asturian: defer
> $ Balkar: defer
> $ Chechen: Accept into ISO 639-1; the code: "ce"
> $ Dargwa: defer
> $ Erzya Mordvin: defer
> $ Franco-Provencal: defer
> $ Gagauz: defer
> $ German, Low; Low German: defer (Already discussed for ISO 639-2)
> $ Inari Sami; Sami, Inari: defer
> $ Ingush: defer
> $ Kabardian: defer
> $ Kalmyk: defer
> $ Karachay : defer (same as Balkar)
> $ Karaim: defer
> $ Karelian, North; North Karelian: defer
> $ Kildin Sami; Sami, Kildin: defer
> $ Kumyk: defer
> $ Lak: defer
> $ Lezghian: defer
> $ Lule Sami; Sami, Lule: defer
> $ Mari, Meadow; Meadow Mari; Mountain Mari: defer
> $ Moksha Mordvin: defer
> $ Nenets: defer
> $ Nogai; Noghay: defer
> $ Provencal: Same as Occitan (post 1500); already in ISO 639-1 as "oc"
> $ Skolt Sami; Sami, Skolt: defer
> $ Southern Sami; Sami, Southern: defer
> $ Tabasaran: defer
> $ Udmurt: defer
> $ Valencian: defer
> $ Walloon: defer

4.3.2 Other languages

> $ Cherokee: defer
> $ Middle Persian; Persian, Middle: defer
> $ Nama: defer
> $ Yi: defer

John Clews

22 March 2000

END OF REPORT TO ISO/IEC JTC1/SC22/WG20