Naming all those languages

Arthur Reutenauer

Royal Opera House, Covent Garden, London

1 May 2015, BachoTEX

Language codes

What's in a language name? Well, two letters, apparently — if you're Very Important.
The International Organisation for Standardisation (ISO) defines, in its standard ISO 639-1 (*Codes for the representation of names of languages, part 1*), two-letter codes for about 200 languages.
Examples: `en`, `pl`, `nl`, `es`, `sk`, `sl`, `sv`, `fr`, `ko`, `ja`.

More language codes

ISO 639-2 defines *three*-letter codes for about 400 languages, including the ones defined by ISO 639-1. You still have to be Important.
It may have two different codes for a language, one *bibliographic* (B) and one *terminological* (T). Examples: `eng`, `pol`, `dut` / `nld`, `spa`, `slo` / `slk`, `slv`, `swe`, `fre` / `fra`.
It also defines codes for *family* of languages: `gem` (Germanic), `sla` (Slavonic).
The list is maintained by the U. S. Library of Congress, called the registration authority.

Even more language codes

ISO 639-3 defines three-letter code for *all* the languages, about 7000 altogether. That includes ancient languages (Sanskrit, Latin) and historic languages such as Middle English and Old English, when they are documented. It is identical with ISO 639-2/T for languages that are included there.
A very interesting development are *macrolanguages*: languages that are also groups of languages.
The Registration Authority is SIL International, whose role is somewhat controversial. Discussions about what is a language or a dialect usually go wrong very soon.
Interesting to note is that all parts of ISO 639 defines codes for *names* of languages.

Even even more

There are more parts to ISO 639: part 4 defines processes and general principles, and part 5 extends the collection part of part 2; it is also maintained by LOC.

No more codes (yet)

Finally, ISO 639-6 was an attempt to produce *four*-letter codes for all languages and wanted to define a full classification of languages. It was never completed and has been withdrawn.

Countries

Some other codes! ISO 3166-1 defines three sets of code for countries, two-letter (alpha-2), three-letter (alpha-3) and numerical (3 digits).

Examples

- GB / GBR / 826
- US / USA / 840
- PL / POL / 616
- NL / NLD / 528
- ES / ESP / 724
- MX / MEX / 484
- SK / SVK / 703
- SI / SVN / 705
- SE / SWE / 752
- FR / FRA / 250
- CA / CDN / 124
- KR / KOR / 410
- JP / JPN / 392

More on country codes

The alpha-2 codes are *nearly* identical to TLD (see `GB` / `.uk`). They are generally *not*, however, identical to ISO 639-1 for languages, even if the names of a language may seem to coincide with the name of a country. My favourite examples: `sl` and `SI`, `sv` and `SE`, `ko` and `KR`, `ja` and `JP`.

There is also some confusion with a few codes, such as `CS` that was registered for Czechoslovakia, then retired, then reused *again* for Serbia and Montenegro (*Srbija i Crna Gora*), then finally retired — and now it will not be reused.

A piece of evidence

Where is this car from?

Scripts

Writing systems (scripts) also have codes, four letters long, in the ISO 15924 standard. Example: `Latn`, `Cyrl`, `Hani`, `Runr`.
The registration authority is the Unicode Consortium.

What about typography?

Think we've got enough codes already? Microsoft doesn't. As part of the OpenType specification, a *different*, ***incompatible*** set of language codes has been defined. There are three letters long, except when they're four letters long.

This makes me sad. But at least there's an equivalence table.

OpenType also defines script codes, that *are* identical, thankfully. Well, almost.

And more

None of these standards and lists is enough anyway.
How to tag British English as opposed to American English? German in the “old” or ”new” spelling? Traditional and Simplified Chinese? Polytonic and Monotonic Greek?
We need a way to combine all these different codes, and also define additional elements in some cases.

The solution

Fortunately, there is such a standard, and it's no messier than the combination of all the previously named standards. It's defined by the Internet Engineering Task Force (IETF) and is publish as a set of Requests for Comments (RFC), (currently RFC 5646 and RFC 4647). Since the RFC numbers change as the standard evolves, it also has a fixed number: BCP 47 (for *Best Current Practice*).

More on BCP 47

It uses all of the ISO standards quoted above, and its own set of subtags maintained in a separate registry.
It can have (amongst others):

- A language code — 2 or 3 letters, whatever's shorter
- A script code
- A country code
- Additional stuff

Subtags are never deprecated. Private use elements can be appended after the prefix `x-`
Thus: `en-GB`, `en-US`, `de-1901`, `de-1996`, `zh-Hant`, `zh-Hans`, `el-monoton`, `el-polyton`.
Actual uses: hyph-utf8. Also Polyglossia, One Day™.

A plea to anyone using language tags

Use the LSR!

`http://www.iana.org/assignments/language-subtag-registry`

The website `http://www.langtag.net/` is also a good resource.

And if you think that's too complicated, *go make a violin*! ← Joke.