Standards for language coding: the ISO 639 family

Rebecca Guenther

Library of Congress

Jan. 8, 2010

ISO Standards development

- **ISO consists of Technical Committees (TC) with subcommittees (SC)**

- **ISO language coding standards are maintained by**
  - TC 37/SC2 (Terminology and other language and content resources)
  - TC 46/SC4 (Information and documentation)

ISO 639 standards

- ISO 639-1: 2-character codes (136 codes)
- ISO 639-2: 3-character codes (450+)
- ISO 639-3: 3-character codes (7700+)
- ISO 639-4: principles
- ISO 639-5: 3-character codes (114)
- ISO 639-6: 4-character codes (??)

ISO 639 Joint Advisory Committee

- Established to advise the RAs for ISO 639-1 and ISO 639-2

- Rotating chairs: Infoterm (for TC37) and Library of Congress (for TC46)

- Committee consists of 3 members of each TC, representatives of each registration authority and up to 6 observers

- Coordinates development of different parts of ISO 639

ISO 639 language coding principles

- Language codes are not changed for stability of standard
- If a language code is retired it is not reassigned to something else
- Programming languages are not in scope
- Only deals with languages; codes from other ISO standards may be added as needed for more granularity, e.g. country codes, script codes

ISO 639-1

- First published 1967
- Covers major languages of the world
- Alpha-2 codes; only 676 possible combinations
- Developed for use in terminology applications
- Consists of a subset of ISO 639-2 and ISO 639-3
- No new 639-1 codes are added if a 639-2 code already exists
- Infoterm is Registration Authority

ISO 639-2

- First published 1998
- Nine years in development by Joint Working Group
- Compromises resulted in 20 alternative codes
- Alpha-3 allows for more combinations than alpha-2
- Based on a widely used bibliographic standard
- Includes individual and group languages
- New requests must satisfy requirements for individual coding
- Emphasis on written languages
- Includes living, ancient and constructed languages
- Library of Congress is maintenance agency

ISO 639-2 criteria

- Evidence of at least 50 documents
- Size and variety of literature
- National or regional support
- Formal or official status
- Formal education
- Other considerations
  - Script
  - Orthography
  - Dialects
  - Group languages

ISO 639-2 approval process

- Requests must satisfy established criteria and a form filled out
- ISO 639-1 codes are not added unless it is an entirely new language to be added
- Committee follows rules for creation of codes as in ISO 639-2 normative text
- Needs unanimous ballot; if not second vote must result in 5 votes to pass

ISO 639-3

- A complete enumeration of all known individual human languages
- Living languages derived from Ethnologue
- Additional extinct, ancient, historic, and constructed languages from the Linguist list
- Does NOT include group languages
- Establishment of 639-3 has resulted in fewer additions to 639-2
- Same rules about scripts, dialects and orthographies as ISO 639-2
- SIL is Registration Authority
- http://www.sil.org/iso639-3/

Impact of ISO 639-3 development on 639-2

- Concept of "macrolanguage": many languages in 639-1 and -2 correspond in a one-to-many manner with individual languages in 639-3
- ISO 639-3 is a superset of the *individual* languages in 639-2
- Group languages are also coded in 639-5
- Many ambiguities of 639-2 were resolved in development of 639-3

ISO 639-3 approval process

- Updated versions released once a year
- Names of languages may be changed
- Dialects are not given separate code elements
- Denotation of a code element may be broadened but not refined
- Existing code element can be retired and replaced by two code elements if determined that code was too broad
- Code elements may be merged if determined that an established individual language is really a dialect
- Change request index: http://www.sil.org/iso639-3/chg_requests.asp

ISO 639-4

- General principles of language coding and application guidelines
- Relationships between parts of ISO 639
- Maintenance of the code sets
- Combining language identifiers with other standard codes
- Currently in FDIS with comments being considered

ISO 639-5

- Alpha-3 code for language families and groups
- Separates into a separate list the language groups included in ISO 639-2 with additional groups
- Language group codes are used when an individual language is not separately coded in 639-2
- Supports overall language coding in 639 series but not a scientific classification of all languages
- Not intended to be comprehensive
- Library of Congress is Registration Authority
- http://www.loc.gov/standards/iso639-5/

ISO 639-6

- Alpha-4 identifier for language variants
- Establishes a hierarchical framework enabling relationships between language variants, families, and groups
- Complementary to and compatible with other parts of ISO 639
- Most specific of the ISO 639 standards
- Recently approved; website under development
- GeoLang Ltd is Registration Authority

IETF Language tags

- RFC 5646 and RFC 4646
- Used in computing standards
- Uses the ISO language coding standards with optional subtags
- Gives a mechanism to combine different standards (e.g. script, region subtags)
- Establishes a subtag registry for language variants maintained by IANA
- Now incorporates 639-3 and 639-5

Developments in maintenance of code lists

- **ISO "concept database" to become master of all in 639 series**

- **Library of Congress is experimenting with a web service for controlled vocabularies**

LC's web service for controlled vocabularies

- Uses semantic web technologies for expressing properties and relationships of language codes
- Uses Simple Knowledge Organization System (SKOS) markup to express these
- Rich information about relationships between concepts (i.e. languages represented by codes)
- Inspired by the linked data movement
- http://id.loc.gov (in future)
- ISO 639-5 data is live using this technology

por

Preferred label(s) by language and script:
    por (Notation)

Used for:
    Portuguese (Language: English; Script: Latin)
    portugais (Language: French; Script: Latin)

Notation label:
    por

Exact matches from other ConceptSchemes:
    http://www.loc.gov/standards/registry/vocabulary/iso639-1/pt (ISO 639-1 Codes or the Representation of Names of
    Languages -- Part 1: Alpha-2 code)
    http://www.loc.gov/standards/registry/vocabulary/languages/por (MARC Code List for Languages)
    http://www.sil.org/iso639-3/documentation.asp?id=por (Codes for the representation of names of languages - Part 3: Alpha-3
    code for comprehensive coverage of languages)

Definition(s):
    This Concept has not yet been defined.

Last update:
    2006-07-19T08:41:54.000-05:00

Concept scheme membership:
    ISO 639-2 Codes for the Representation of Names of Languages -- Part 2: Alpha-3 code

Collection membership:
    ISO 639-2 Collection
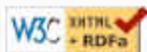   ISO 639-2 Bibliographic Codes Collection

Concept status:
    stable

URI for this Concept:
    http://www.loc.gov/standards/registry/vocabulary/iso639-2/por

This resource is an instance of:
    Resource Description Framework: Resource
    Simple Knowledge Organization System: Concept

ISO 639-2 language code in SKOS

```xml
<rdf:Description rdf:about= "http://www.loc.gov/standards/registry/vocabulary/
iso639-2/por">
        <rdf:type  rdf:resource="http://www.w3.org/2008/05/skos #Concept"/>
        <skos:prefLabel xml:lang="x-notation">por</skos:prefLabel>
        <skos:altLabel xml:lang="en-Latn">Portuguese</skos:altLabel>
        <skos:altLabel xml:lang="fr-Latn">portugais</skos:altLabel>
        <skos:notation rdf:datatype="xs:string">por</skos:notation>
        <skos:definition xml:lang="en-Latn">This Concept has not yet
been            defined.</skos:definition>
        <skos:inScheme            rdf:resource="http://www.loc.gov/
standards/registry/vocabulary/iso639-2"/>
        <vs:term_status>stable</vs:term_status>
        <skos:historyNote
rdf:datatype="xs:dateTime">2006-07-19T08:41:54.000-        05:00</skos:historyNote>
        <skos:exactMatch rdf:resource=  "http://www.loc.gov/standards/
registry/vocabulary/iso639-1/pt"/>
        <skos:exactMatch rdf:resource=  "http://www.loc.gov/standards/
registry/vocabulary/languages/por"/>
        <skos:changeNote rdf:datatype="xs:dateTime">2008-07-
09T13:49:05.321-04:00</skos:changeNote>
</rdf:Description>
```

Conclusions

- Needs for language coding vary by application, so multiple standards are needed
- There is a high degree of compatibility between the ISO 639 code lists
- Common principles are followed, such as stability of the lists as a high priority
- Centralization of maintenance in the new ISO concept database structure should result in further consistency between the standards