



Improved Language Coding Efforts and Issues

Unicode Conference
January 2002



Language Codes

- Designators for languages, dialects, etc.
 - E.g., ARA or AR for Arabic
 - Used in many forms (e.g., XML, statistics, cataloguing, etc.)
- Used for designation of
 - **Tools** (e.g., spelling checkers, grammar checkers, hyphenation dictionaries, dictionaries, search engines, etc.)
 - **Materials** (e.g., books, documents, paragraphs, abstracts, table of contents, audio, video, librettos, etc.)
 - **People** (e.g., those speaking Chinese at home; those who are offering translation services in a certain language, etc.)
 - **Locales** (e.g., Belgium French market requirements)



Types of Information

- Language
- Dialect
- Geographic Area of Use
- Locale (similar to above)
- Language Family or Group
- Orthography
- Transcription
- Modality
- Time



Situation Now

- Insufficient ISO codes to cover all languages and dialects
- Inconsistency of data definitions
- Inconsistency of linguistic definitions
- Conflicting standards
- Specification of too limited standards for language codes (e.g., Java)
- Little framework



Purpose

- Provide information on efforts regarding language codes
- Discuss requirements, issues, and solutions
- Obtain feedback from you on your applications, requirements, issues, and suggestions



Agenda

- Introduction
- Background
- Present Efforts
- Jennifer DeCamp
 - MITRE, ISO TC 37, OSCAR
- Rebecca Guenther
 - Library of Congress, ISO TC 46
- Håvard Hjulstad
 - ISO TC 37
- Sue Ellen Wright
 - Kent State University, ISO TC 37, OSCAR
- Monty George
 - U.S. Department of Defense
- Peter Constable
 - SIL International
- David Dalby
 - Linguasphere Observatory
- Requirements
- Issues
- Solutions
- Questions and Feedback



ISO 639-2 Development

- Joint working group of ISO TC37/SC2 and ISO TC46/SC4
 - TC37/SC2: Terminology and lexicography
 - TC46/SC4: Information and documentation/Technical interoperability
- Nine years of development, 1989-1998
- Recognized need for a larger list of languages than alpha-2 code
- Based on a well-established language code list



NISO Z39.53 and MARC Code List for Languages

- Used since 1968 by libraries, information centers, indexing services, archives, publishers etc. in large computer systems
- Language codes to indicate
 - Language of resource
 - Language of summary/abstract
 - Language of table of contents
 - Language of accompanying material
 - Language of original for translations
- Used by systems for resource discovery and identification, limiting result sets



ISO 639-2 principles

- Used to identify a language or language group
- Not intended as abbreviations but code used by computers
- Systems can display a language name instead of the code itself
- Not intended to be comprehensive; languages represented have a significant body of literature
- Need for continuity and stability in large databases; codes rarely changed



ISO 639-2 principles

- Collective codes used for languages without sufficient documents to qualify for a separate code
- If written in more than one script assigned only one code
- Dialects represented by language code for major language
- Languages using more than one orthography given only one code



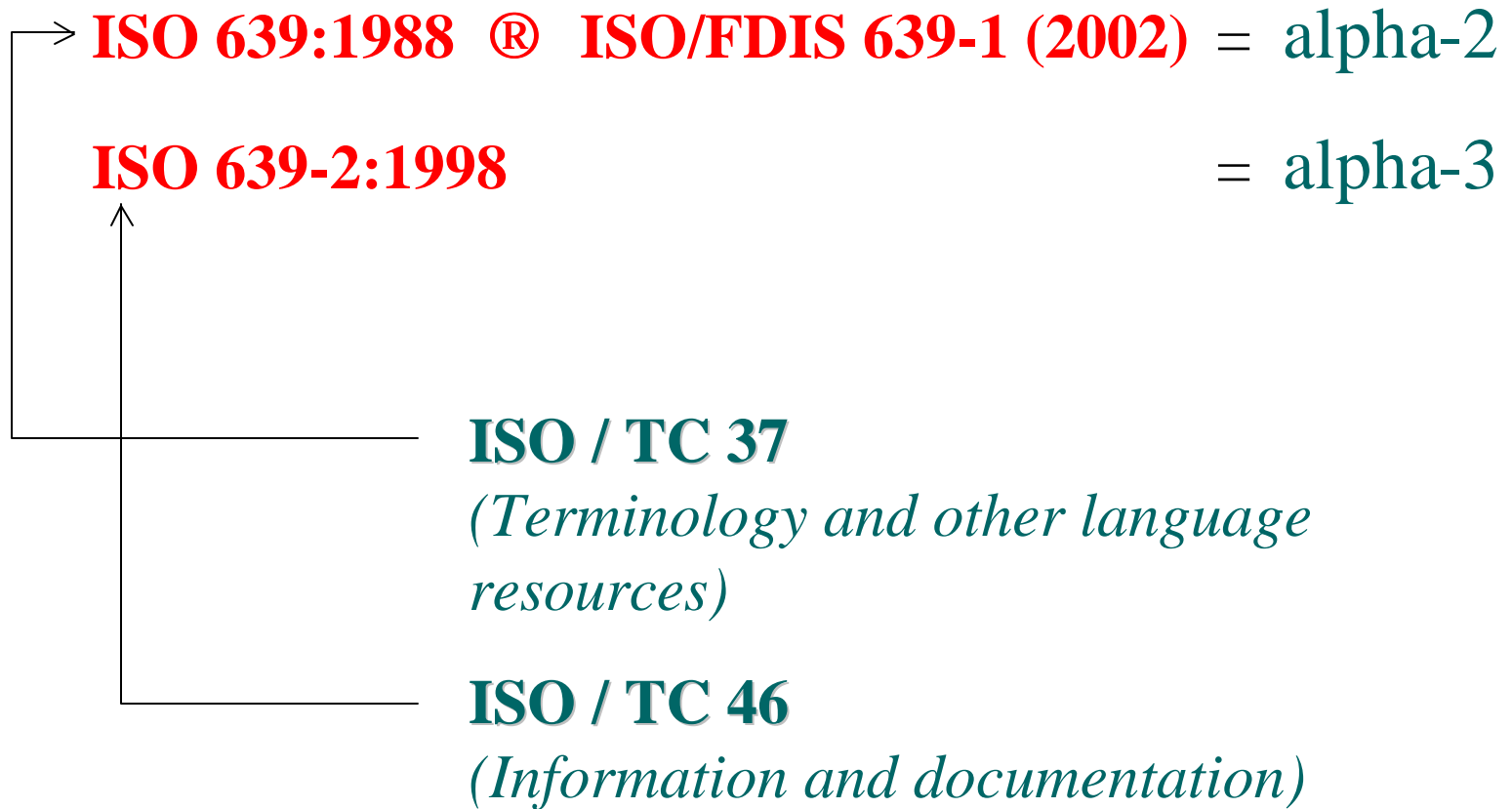
ISO 639 Joint Advisory Committee

- Established 1998 with approval of ISO 639-2
- 3 voting members and up to 3 observers from each TC
- Registration authority initially processes application for new codes; voting by JAC
- Registration authorities:
 - Infoterm for ISO 639-1
 - Library of Congress for ISO 639-2
- Chair rotates between LC and Infoterm



Uses of ISO 639-2

- Libraries and information centers with millions of bibliographic records
 - 12 million in LC; 46 million in OCLC
- Emerging metadata applications
 - Dublin Core Metadata Initiative
 - ONIX (publishers)
- Resource discovery and identification that requires less granularity than other applications





Background, history of ISO 639

ISO 639 has been shaped by the needs of

⇒ documentation, libraries, bibliography

⇒ terminology and lexicography

⇒ language resources and language technology

The needs are different!



Maintenance of ISO 639-1 and 639-2

ISO 639-1 Registration Authority
(Infoterm, Vienna)

ISO 639-2 Registration Authority
(Library of Congress, Washington DC)

Joint Advisory Committee (JAC)



“Development of ISO 639-1 and ISO 639-2 will remain conservative”

**i.e.: ISO 639-1 and ISO 639-2
will not meet the users’
requirements as to granularity
and coverage**



What does ISO / TC 37 want to do?

- **Work within and outside alpha-2 and alpha-3**
- **Define a model for language identification**
- **(Attempt to) define “language”**
- **Develop specifications for “modifiers”**
- **Specify default values for modifiers within languages**
- **“Mass encoding”**
- **Hierarchical identifiers**



What may our users expect from TC 37?

- **variable length identifiers** ☹️
- **improved coverage** 😊
- **synonyms** ☹️
- **hierarchy identifiers, group identifiers** 😊
- **variant coding mechanisms** 😊



Interactive Standards

- ISO TC 37/SC 2: ISO 639-1 and 639-2; potential for alpha 4, extensions of 639-2
- IETF RFC 3066 (based on 639)
 - Obsoletes RFC 1766
 - 639: alpha 2, then alpha 3, no synonyms
- W3C Recommendations: xml:lang
- Unicode
- JTC 1/SC 22/WG 20: Locale codes: language codes + non-linguistic information
 - Specification Methods for Cultural Conventions



Applications for Various Standards

- Different programming environments require different attributes
 - lang, xml:lang, locale identifiers at different levels in the same resource
- Format constraints
 - XML rules for specified attribute values & targets
- Identification of code components
- Semantics for combining components to produce codes that meet the needs of different environments



Requirements

- Scope
 - for US Government



Requirements

- Example usages
 - tools
 - buy / make
 - functionality
 - resource management
 - justification
 - people
 - who? what? how?



Requirements

- Themes

- commercial products / interoperability
- language codes
- related coding needs
 - writing systems
 - orthographies...
- intertwined needs
 - tools, people, codes
- comprehensive solutions needed



Challenges

- Constable & Simons 2000
 - theoretical & practical issues
 - definition(s) of language
 - other categories
 - meanings of codes
 - *SIL Ethnologue* offers potentially significant solutions
- Constable & Simons 2002—bring order to existing implementations



Challenges

- need to make better sense of user “requirements”
 - e.g. “US dialect” of English
 - e.g. *How do I indicate Simplified Chinese?*
 - need to identify the types of language-related category needed
 - need to arrive at operational definitions
 - need to assess usage scenarios for which each type is relevant



Challenges

- need ontological model for languages and language-related categories that guides the formulation of solutions
- need to develop comprehensive solutions
 - cover full range of needs
 - sensible structure guided by model
 - accommodate existing implementations



Questions and possible solutions

- What is an optimum “language” for coding purposes ?
- What is a valid “language group” for coding purposes ?
- What is the best structure for basic language codes ?
- What is best procedure for expanding ISO 639 ?



Questions and possible solutions

- What place would the existing 2- and 3-letter codes of ISO 639 have alongside an expanded set ?
- What set procedures should be envisaged for the coding of varieties within languages ?
- How might language codes, whether extended or not, be conveniently distinguished from other alphabetic sequences, and also classified?



Questions and possible solutions

- A selection of possible solutions is incorporated in preparation of Linguasphere-2
 - 2nd edition of Linguasphere Register of World's Languages and Speech Communities (2003)