

Building Supporting Structures

Lee Gillam
WLDC
&
University of Surrey

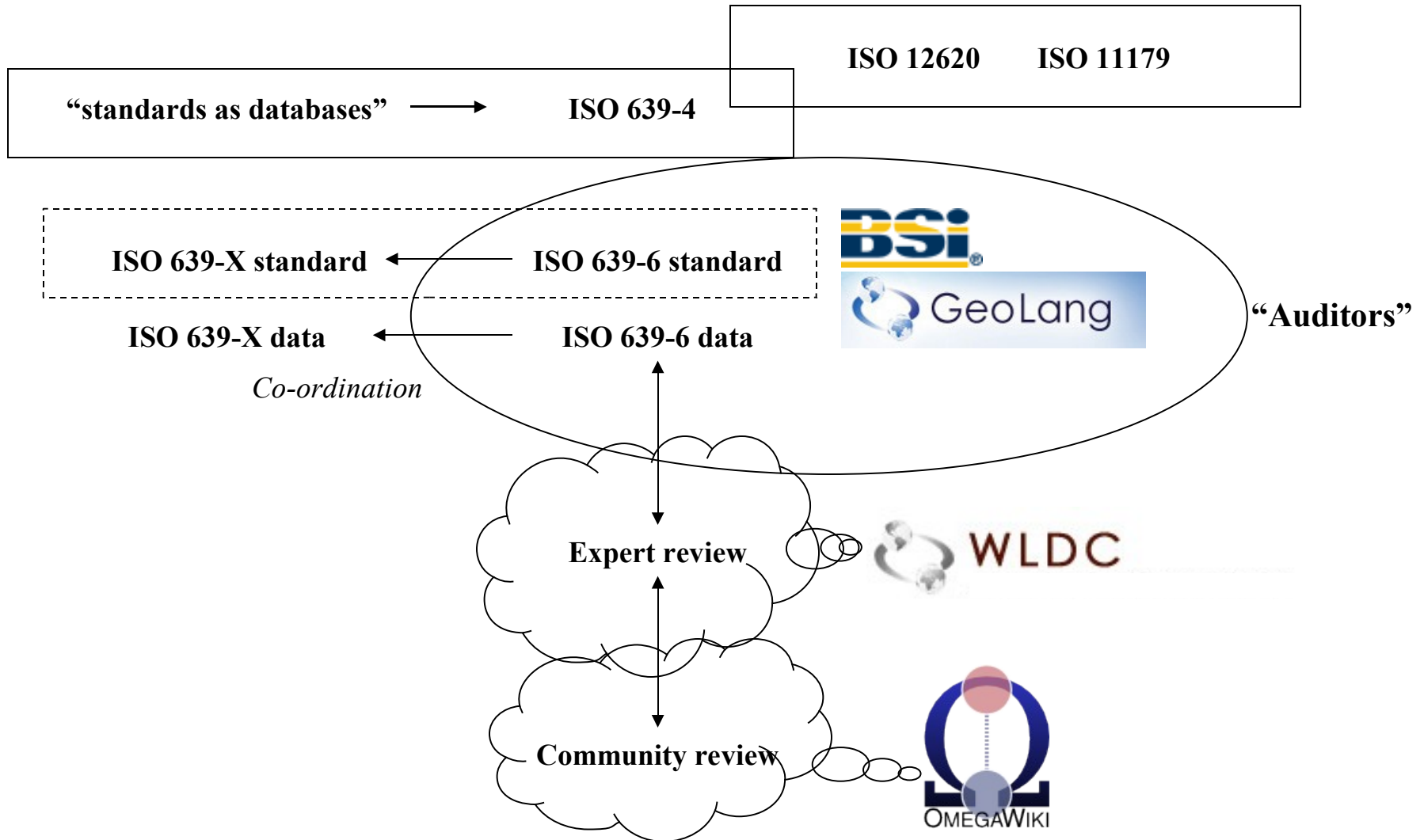
Overview

- Language Codes Standards are growing in number and complexity
 - From 2 to 6
 - From 400 identifiers to upwards of 30000
 - From lists to databases
 - From tables to metadata registries
 - From published text documents to “published” databases
 - From IETF RFC to RFCs to RFCs
 - From a closed membership committee to an open Community initiative (OmegaWiki)
 - with accompanying (web) services and products

Overview

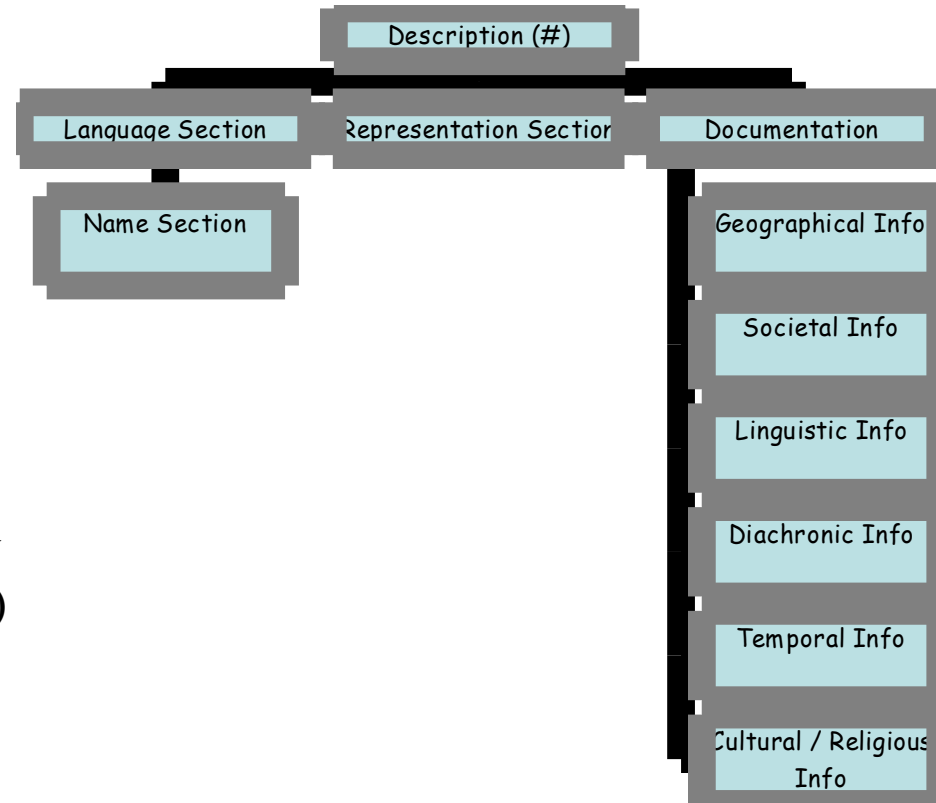
- Language Codes Standards are growing in number and complexity
 - From 2 to 6 – eventually back to 1?
 - From 400 identifiers to upwards of 30000 – plus supporting metadata
 - From lists to databases – multiple metadata registers
 - From tables to metadata registries – registers + policies + “auditors”
 - From published text documents to “published” databases – “SAD”
 - From IETF RFC to RFCs to RFCs – consume, consume, consume
 - From a closed membership committee to an open Community initiative (OmegaWiki) – supporting infrastructure, expert review of community contributions (e-Voting?)
 - with accompanying (web) services and products – Open Source and bespoke, and secured funding as necessary

Overview



The Language Documentation and Interchange Format (LDIF)

- Language Documentation associated to metadata – interoperability with DCIF (12620) as a “subset” of LDIF (LDF: 639-4 s9)
- Notions of “Identifiers” – reference name vs alpha-2/3/4.
 - Names => ML thesaurus; representations as governed by rules in standards (alpha-2/3/4)
- Document => Representative (c/w country flags)



The Language Documentation and Interchange Format (LDIF)

9.2.2.3 Temporal Information

- historical note;
- modern events and changes;
- date.

9.2.2.4 Diachronic Information

For recording changes between versions:

- historic class;
- historical classification.

9.2.2.5 Cultural and religious Information

- community;
- religious culture.

9.2.2.6 Societal Information

- population size;
- social status;
- legal status;
- speaker identification;
- migration;

9.3.3 Country information

- official country name: name used by the country in its official documents;
- country population: approximate number of people in a country;
- geographic reference;
- national language: language spoken by a large portion of the population of a nation;
- official language: language designated by an official body;

9.3.5 Reference materials

- source (“literature”);
- dictionary;
- grammar;
- broadcast media;
- braille literature.

- country literacy rates: estimate of the percentage of the population in the country that is literate in some language;
- non-indigenous language [or immigrant language]: language spoken in one country by a community that has migrated from another country where there is no significant dialect difference between the two locations.

9.3.6 Geographical information

- geological information = [altitude range or ...] (physical setting of the society);
- altitude range;
- ecological information = [subsistence type or ...] (general economic adaptation of the society).

9.3.7 Sociocultural information

- religion: religious affiliation of people.

Language Codes & Language Resources

ISO 16642 TMF	ISO 24611 MAF	ISO 24612 LAF	ISO 24613 LMF	ISO 24615 SynAF	ISO nnnnn SemAF
Terminological, Morphosyntactic, Linguistic, Lexical, Syntactic and Semantic Data Categories (Metadata)					
ISO 12620: Data Categories					
ISO 639-1-6 (x4) Language Codes					
ISO 639-4: Language Code metamodel					
ISO 11179: Metadata Registries					

Applications in e-Science, video annotation and document quality

LMF: Lexical-Semantic and Ontological Resources workshop

Summary of Parts

Title of Standard	Status	Registration Authority	Number of identifiers (approx)
ISO 639-1: Part 1: Alpha-2 code	Published (2002)	InfoTerm	150
ISO 639-2: Part 2: Alpha-3 code	Published (1998)	Library of Congress (LoC)	400
ISO 639-3: Part 3: Alpha-3 code for comprehensive coverage of languages	Published (2007)	Summer Institute of Linguistics (SIL)	7000
ISO 639-4: Part 4: Implementation guidelines and general principles for language coding	Expected late 2007.	n/a	n/a
ISO 639-5: Part 5: Alpha-3 code for language families and groups	Expected late 2007.	TBC	100
ISO 639-6: Part 6: Alpha-4 representation for comprehensive coverage of language variation	Expected early 2008.	GeoLang	25000