Language Codes Standards

Lee Gillam
University of Surrey

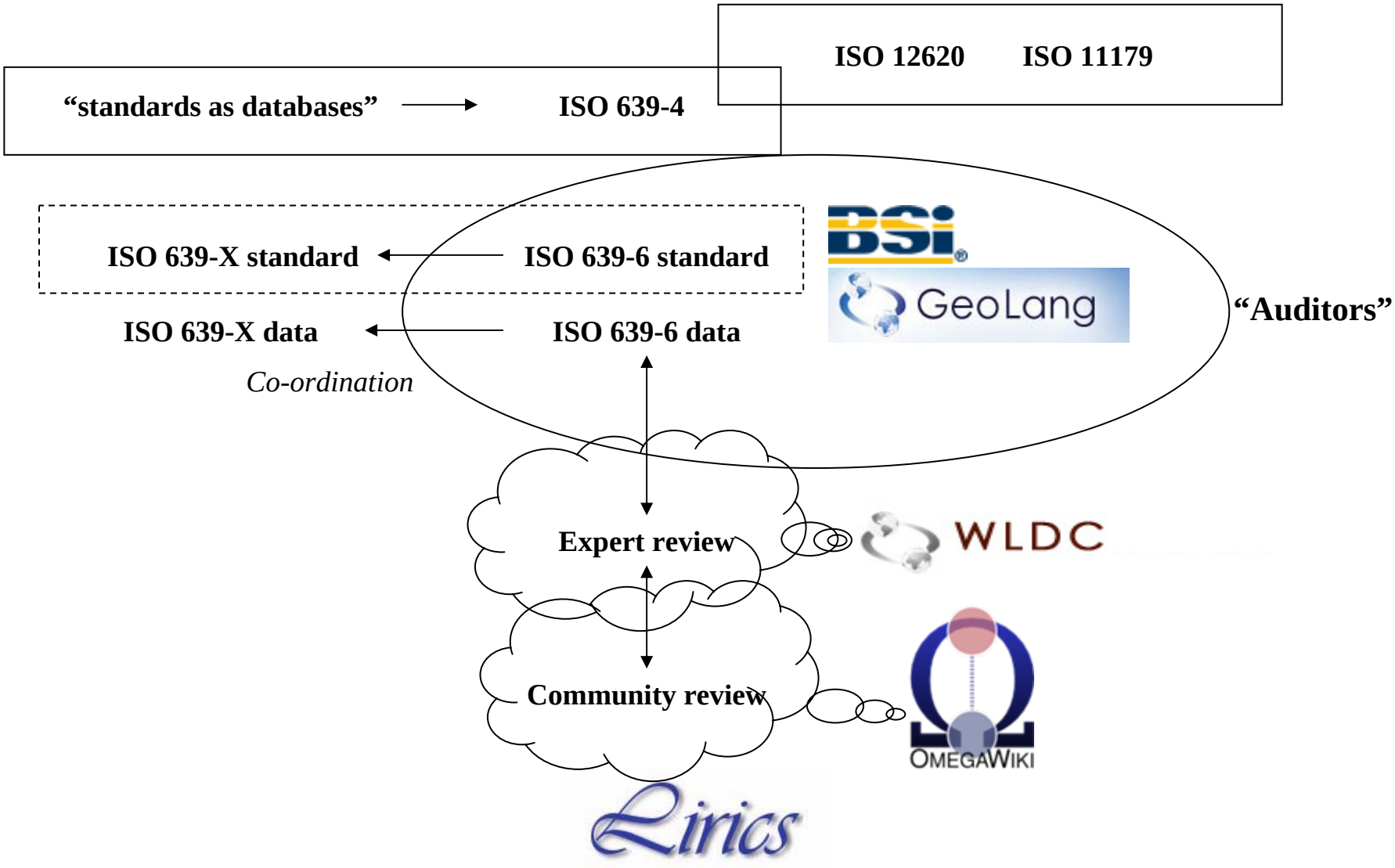
Overview

- 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


ISO 12620
ISO 11179
"standards as databases"
ISO 639-4
ISO 639-X standard
ISO 639-6 standard
ISO 639-X data
ISO 639-6 data
Co-ordination
"Auditors"
Expert review
Community review
WLDC
OMEGAWIKI
Lirics

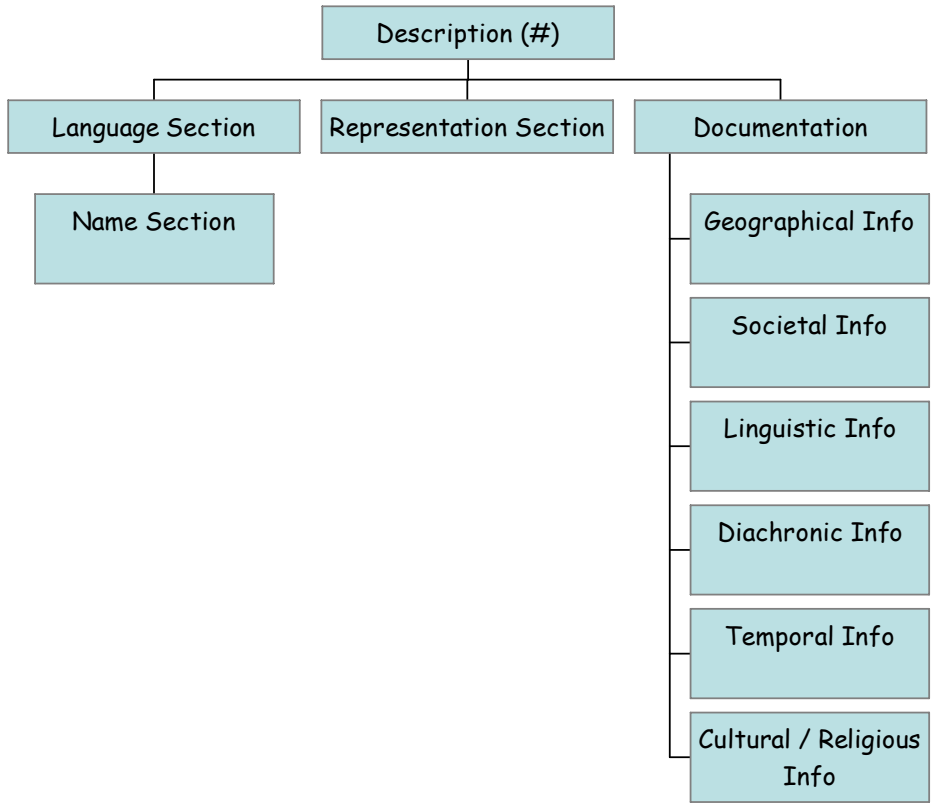
Language Codes & Language Resources

ISO 16642 TMF	ISO 24611 MAF	ISO 24612 LAF	ISO 24613 LMF	ISO 24615 SynAF	ISO nnnnn SemAF
Terminological, Morphsyntactic, 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…..

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;

— 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.

9.3.5 Reference materials

— source ("literature");

— dictionary;

— grammar;

— broadcast media;

— braille literature.

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.

Shapes of Things to Come

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*

Summary

- Language experts may identify linguistic content in a highly precise manner – what can non-experts do? How can we assist this process?
  - Providing more specific sets of labels may help in discovery of written or spoken languages in all kinds of media – and help to harmonize research activities - so long as people know what they are looking at.
  - Inaccuracies of currently tagged content; need to take the responsibility away from end users?
- More precise identification improves the chances of getting the right thing – consider “coffee” vs. “coffee + TYPE + COLOUR …” vs. “strong black coffee, in a mug, with 2 sugars”.
- Beyond documentation of names and representations, documentary information for each language might be helpful for – and for opportunities beyond those we might consider today.
  - Working towards a machine-readable representation for all such information is a longer-term goal.

Motivations

- **Scientific:** Repeatability of science: to examine others' work
  - Need to be able to use near-identical tools, techniques, and data
  - Data at the *exascale* - a few new exabytes (billion GB) of unique information / year
  - **Sciences need basic standards** – e.g. SI units; periodic table, languages
- **Socio-Economic:** Addressing the "digital divide"
  - The "recognition of the regional or minority languages as an expression of cultural wealth" (European Charter for Regional or Minority Languages); The World Summit for the Information Society's (WSIS) drive towards an Information Society for All and UNESCO's multilingualism and literacy for all programmes:
    - "860 million adults are illiterate, over 100 million children have no access to school, and countless children, youth and adults who attend school or other education programmes fall short of the required level to be considered literate in today´s complex world" (UNESCO)
  - Facilitating transfer of local and indigenous knowledge, e.g. in African countries, is important for reasons of identity, ecology, agriculture, and social and cultural maintenance.
- **Preservation:** Increases required in ability to **manage** and **analyse** collections of language.

Motivations

- **Technological:** New Media and ever-larger volumes of information:
  - **Written:** 24-hour news coverage published on news-wires and more recently using Really Simple Syndication (RSS) feeds; MySpace; Flickr
  - **Spoken:** digital radio and on-demand podcasts; continuous televisual coverage through the variety of available digital television channels; broadband television efforts such as IET.tv, and other forms of IPTV and on-demand programme schedules referred to as "timeshift TV"; CCTV; YouTube
- **Discovery and collation:** New, and old, collections of spoken languages:
  - BBC Voices of the British Isles – "from Shetland to Penzance. Eavesdrop on Rotarians in Pitlochry and Travellers in Belfast. Drop in on skateboarders in Milton Keynes. Overhear pigeon fanciers in Durham." http://www.bbc.co.uk/voices/
  - Ancestral voices: collection at Humboldt University in Berlin - "An Austrian scholar was fascinated by the diversity of English dialects. So on primitive technology he recorded voices from every English county reciting the Parable of the Prodigal Son". British prisoners of war (WW I) reciting the parable of the prodigal son as the First World War raged around them. *"When he had spent everything a great famine came o'er t' country and he begun to be in want."*

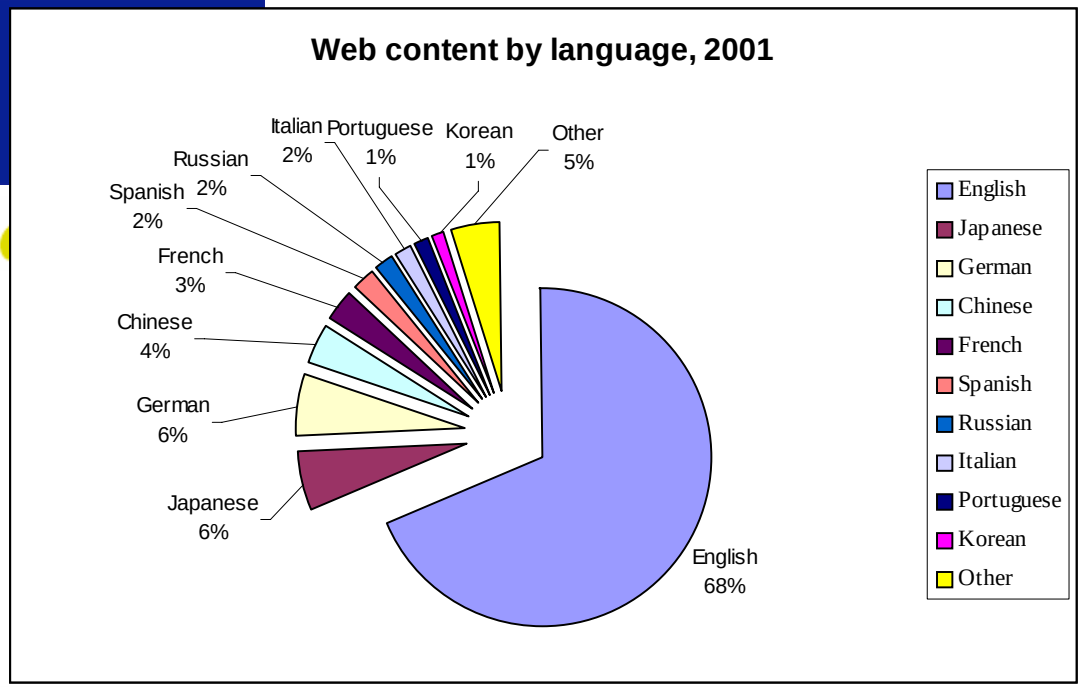
Motivations

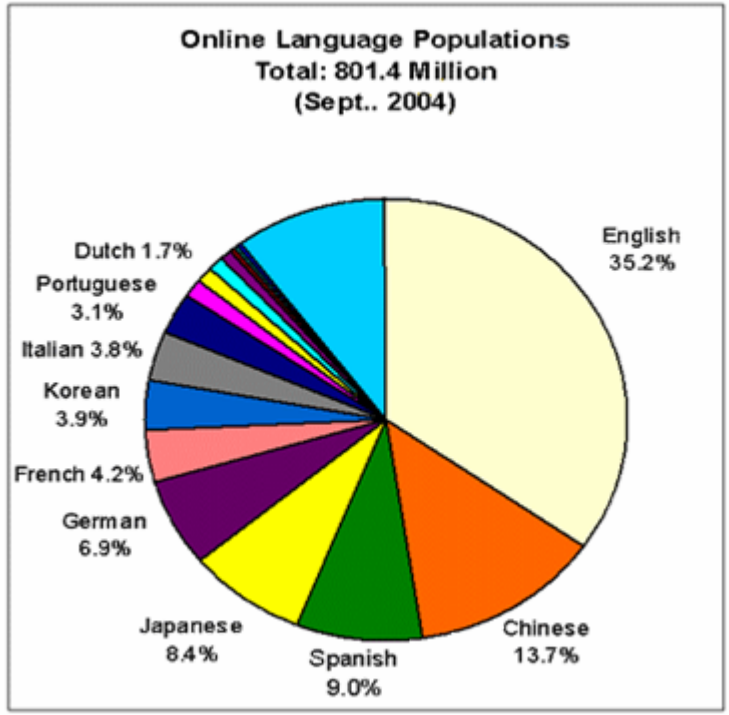
- **The Multilingual Internet?** - September 2005, ICANN approved the first "ccTLD" for a particular human language and culture: '.cat' (Catalan); March 2007, ICANN discussing IDNs (Lisbon)
- **The Semantic Web:** Increased coverage for XML content: XML meant for machines and machines like precision
- **Standardized metadata**
  - Enable convergence of formats, systems and collections to improve reusability / interoperability
  - Systemic efforts across ISO TC37 relating to Language Resources
    - Standards-conformant metadata can improve management of data and facilitate scaling up of research.
    - Research results identified using finer-grained metadata become available for use beyond that originally intended, ethical considerations allowing.
  - BUT: Takes effort (= money) to maintain
  - Efforts for languages comparable to Universal Decimal Classification (UDC)
- **To foster <u>Data Wombs:</u>**
  - Data Tombs: large quantities of data "deposited to merely rest in peace, since in all likelihood it will never be accessed again" (Fayyad and Uthurusamy, 2002)
  - Data Wombs: Living archives "where data collections that will evolve over time are sustained and nurtured" – e.g. news.

Increases in complexity

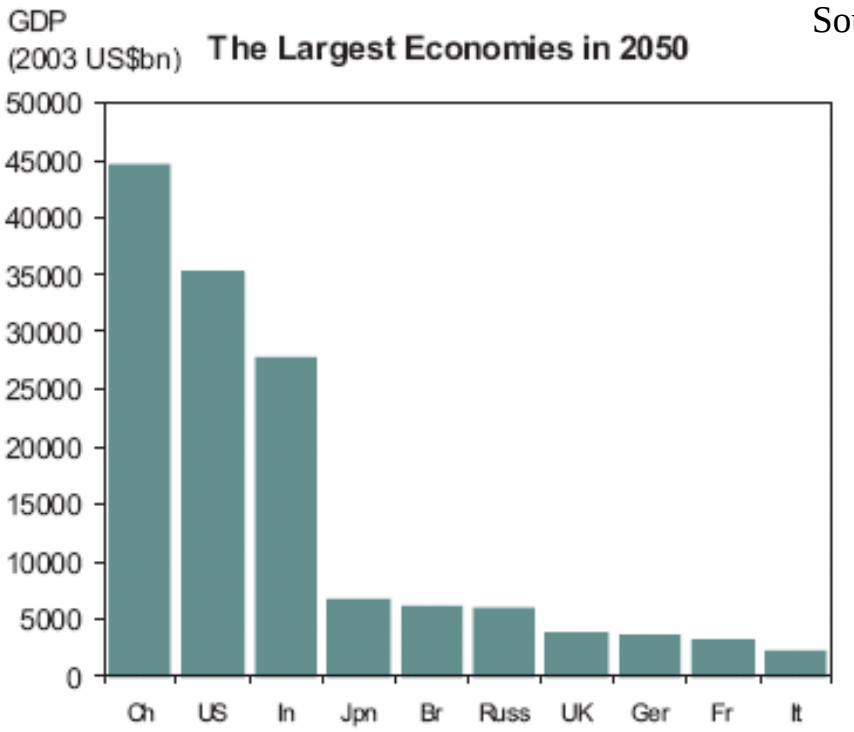
- **Web (*1991*):** Information sharing for people to analyse
- **Semantic Web (*1998*):** The machine-processable web
- **Grid Computing (*1999*):** "An infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources." (Foster/ Kesselmann) – notion of "Virtual Organizations" (VO).
- **Semantic Grid (*2001*):** the machine-understandable Grid – automatic SOAs?
- **Knowledge Grid (*2001*):** human-understandable, Semantic Grid? Knowledge discovery in "data tombs".
  - Build, store, share and use KM techniques for analysing large data sets: to make scientific discoveries? Deriving/inferring new knowledge. Metadata is a key element.
- **Web 2.0 (2006); Second life - Web X.0?? (*2006*+)**

- Standards for:
  - quantities (ISO 31-0); language codes (ISO 639 series); country codes (ISO 3166); dates (ISO 8601); character sets (ISO 10646) and data types (ISO 11404).
- Increasing volumes of data demand increasingly granular metadata – reduction of the information load, or "how to put the problem elsewhere"
  - ISO 639-1 [c. 150] & ISO 639-2 [c. 400]
  - ISO 639-3 [7000+]
  - ISO/DIS 639-5 [(language families)]
  - ISO/DIS 639-6 [25000+]
- Sufficient for today….?





Sources: Various



ISO 639 and related standards

- Storage, management and retrieval of ever-increasing volumes of electronic data, generated in the spoken, written and signed languages.
- ISO 639 parts 1 and 2 are lists of identifiers and names: burden of interpretation and use is on the user.
- Relationships with other items within these lists are unspecified: identifiers are discontinuous and nothing should be inferred from them beyond the fact that those using the same identifier for some content may find content elsewhere that is also marked with this identifier.
- ISO 639 used by the Internet Engineering Task Force (IETF). RFC 1766 -> 3066 -> 4545/6/7 -> ? : Geographical variation by country codes (ISO 3166); Script variation by script codes (ISO 15924); Orthographic variation by request. Common items provide for some degrees of matching.
  - RFCs in turn consumed by XML.
  - Generative: some kind of coffee, in some sort of mug, possibly with sugar

ISO 639 and related standards

- Computer program developers who are used to the production and "inspection" of markup languages will be familiar with e.g. "en-US" denoting "American English". Terminologists will be familiar with assigning a language to the terms for their concepts when working with more than one language.
- Burden of interpretation:
  - With hundreds and thousands of codes, and combinations, to choose from, users should assume nothing:
    - Naïve users, or lazy users, might assume that AF stands for "African", rather than Afrikaans. Or they may assume it is a two-letter country code for Afghanistan.
    - Country code "CS". Was Czechslovakia, has its own, heavily used, ccTLD that was deleted after 1993. Became ISO 3166-1 code for Serbia and Montenegro. S&M have now split from their union. How should users interpret information tagged with CS? Where is information about broader and narrower interpretation contained? How should retrieval systems be coded to cope with such infrequent but significant changes? At a more general level, such issues are important also for so-called "ontology evolution", and more specifically for management of information in metadata registries.
    - If your destination is Genoa, Italy, beware of baggage handlers inferring from the GOA tag that your bags should be sent to India!

ISO 639 and related standards

- For high-level simplistic differentiation, e.g. selecting a spelling checker, these identifiers may be sufficient.
- The 400 or so 639-1/2 identifiers stand for relatively “well-used” languages, and as a consequence there is likely to be a greater wealth of content in these languages. Google: +the = 19Bn +- 12Bn pages; +the +of = 18Bn +- 12 Bn pages. Limited differentiation beyond this.
- Ever-larger digital collections, broad identification becomes of limited value; increased granularity of identification can only help in the organization and management of such collections.
- Larger systems of identifiers need to be managed, not least to try to reduce the burden of interpretation on the users. Languages and their identifiers should be treated as a system of interdependent and related concepts.
  - Changing the definition of a concept necessitates considerations of changes for directly related items – especially superordinate and subordinates
  - Diachronic variation: “atom” c1900 vs “atom” c2000; “CS” => “CS” => “CS”?

Metadata registries

- Metadata registry according to ISO 11179 series of standards, to be in conformity with ISO 12620.
- According to ISO 11179:
  - A **Value Domain** is associated with a **Conceptual Domain**: A Value Domain provides a representation for the Conceptual Domain.
  - Example Conceptual Domain and set of Value Domains is ISO 3166, Codes for the representation of names of countries.
  - **ISO 3166 describes the set of seven Value Domains**: ***short name in English***, official name in English, short name in French, official name in French, alpha-2 code, alpha-3 code, and numeric code.
  - Each representation contains **a set of values that may be used in the value domain associated** with the DEC; each one of the seven **associations** is a data element.
  - For each representation of the data, the **permissible values**, the **datatype**, the **representation class**, and possibly the **units of measure**, are altered.

Conceptual domain name: Countries of the world

Conceptual domain definition: Lists of current countries of the world represented as names or codes.

Value domain name (1): Country codes – 2 character alpha

Permissible values:

<AF, The primary geopolitical entity known as *"Democratic Republic of Afghanistan"*>

<AL, The primary geopolitical entity known as "People's Socialist Republic of Albania">

. . .

<ZW, The primary geopolitical entity known as "Republic of Zimbabwe">

Value domain name (2): Country codes – 3 character alpha

Permissible values:

<AFG, The primary geopolitical entity known as "Democratic Republic of Afghanistan">

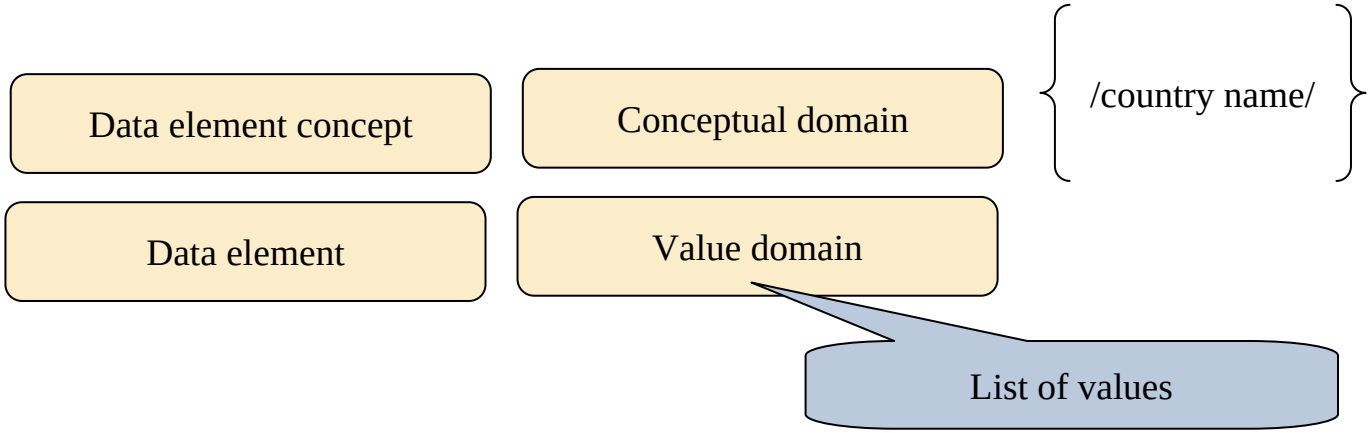
<ALB, The primary geopolitical entity known as "People's Socialist Republic of Albania">

. . .

<ZWE, The primary geopolitical entity known as "Republic of Zimbabwe">

Metadata registries

/Country/



Implemented as an XML *attribute* named ‘country’

country=“?”

<xml country= ’FR ’ >

FR

The Language Documentation and Interchange Format (LDIF)

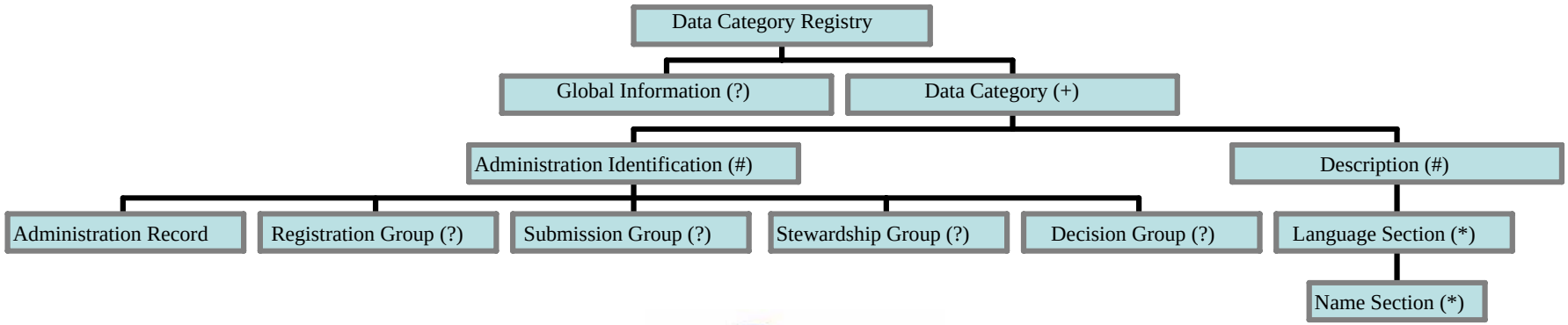
- Model for ISO 639 proposed and developed by LIRICS project participants (Gillam, Romary); recently accepted for inclusion and review in the current iteration of the developing ISO 639 part 4.
  - intended to be fully compatible with models being developed in ISO TC 37 in general, compatible with the Data Category Interchange Format defined in ISO 12620, and to provide a means for interlinking the collection of identifiers provided across the 639 series.
  - ISO TC 37 standards for computational use of terminology collections, specifically ISO 16642 and its combination with ISO 12620, emphasize a ***metamodel*** in combination with metadata identifiers, referred to as ***data categories***.
  - Language identifiers of ISO 639 shall be compatible, interoperable, mutually understandable, and usable to the degree of precision needed by the user up to the limitations of these identifiers.
  - Language identifiers themselves need to be described by metadata.
  - All of these metadata items can be submitted to the metadata registry specified according to ISO 12620

The Language Documentation and Interchange Format (LDIF)

- ISO 639 model based on:
  - need to replicate simplistic structure of ISO 639-1 and 639-2
  - inferred model of the Ethnologue as published
  - emergent model through BSI for ISO 639-6 adapted, generalized and cross-validated from encyclopædic and other sources including:
    - Gordon Jr, R. G (Ed.) (2005). Ethnologue: Languages of the World, 15th Edn. SIL International.
    - Voegelin, C.F. and F.M. (1977) Classification and index of the world's languages. New York, NY: Elsevier North Holland, Inc.
    - Ruhlen, M. (1987) A guide to the world's languages. Vol.1: Classification. London: Edward Arnold.
    - Bernard Comrie (ed.) (1987) The World's major languages. Oxford University Press, New York,
    - Chambers, J.K. and Trudgill, P. (1998) Dialectology. Cambridge: Cambridge University Press
    - Dalby, D (1999). Linguasphere Register of the world's languages and speech communities. Linguasphere Press.
  - development of ISO 639-6 initially assisted by a fund made available by the Department of Trade and Industry of the UK and administered by BSI; subsequent efforts in standardization and validation have been funded, and supported, by BSI and ICT Marketing Ltd.

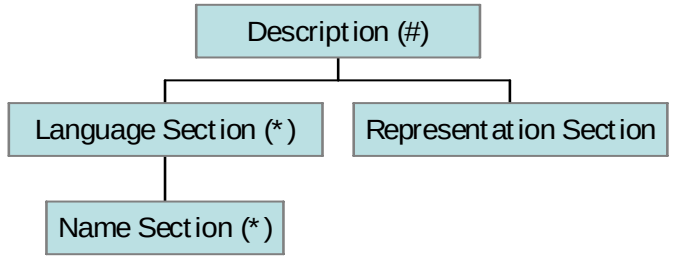
The Language Documentation and Interchange Format (LDIF)

- The 12620 metamodel provides a simple representation for much of the existing ISO 639-1 and 639-2 data.
  - ISO 639 1&2 contain names of languages in both English and French; “Description” in one or more “Language Sections” with one or more names in the “Name Section”.
  - Administrative detail can be filled as needed, but is perhaps less interesting.
- Some variation between the metamodel of the metadata registry and the ISO 639 model described here may be unavoidable, but the intention is that the core of the model be as consistent with ISO 12620 as possible.



The Language Documentation and Interchange Format (LDIF)

- LDIF expands Description: 12620 does not appear to provide for non-linguistic identification – the "identifiers" of ISO 639 – unless one is prepared to 'abuse' Language Section.
  - Register values for all of the different alpha-2, alpha-3, alpha-4, numeric, alphanumeric, hexadecimal, and any other oddly constructed, representations.
  - ISO 639 has multiple "terms for a concept": the "terminological" and "bibliographical" varieties of ISO 639-2 alpha-3s and, for a number of the ISO 639-2 alpha-3s, equivalent ISO 639-1 alpha-2s. LDIF needs to cater, simply, for all of these possibilities. And offer integration beyond this (a + b = c).



[**Representation Section**]
/representation/ {format="alpha-2"; identifier = "gd"}
/representation/ {format="alpha-3"; identifier = "gla"}
[**Language Section**]
/language/ = /en/
[**Name Section**] /name/ = "Gaelic"
[**Name Section**] /name/ = "Scottish Gaelic"
[**Language Section**]
/language/ = /fr/
[**Name Section**] /name/ = "gaélique"
[**Name Section**] /name/ = "gaélique ecossaise"

English (/en/) and French (/fr/) should also be described metadata in the registry, hence available for descriptions of other metadata.

The Language Documentation and Interchange Format (LDIF)

- Use of language identifiers in the documentation of further languages provides potential for a multilingual catalogue of language names (order of 49 million).
  - How to discover the language without knowing one of the names for the language?
- Within the **Description** section, “broader data category” used to associate language identifiers to broader language identifiers. Support for language groups and families, ISO 639-5.
- Consider expansion of the Indo-European languages involving the West Germanic language family (Example data, re-presented, from ISO 639-5)

ALPHA-3	PARENT ALPHA-3	ENGLISH	FRENCH
ine		Indo-European languages	indo-européennes, langues
gem	ine	Germanic languages	germaniques, langues
ine	gem	West Germanic	germanique occidental

The Language Documentation and Interchange Format (LDIF)

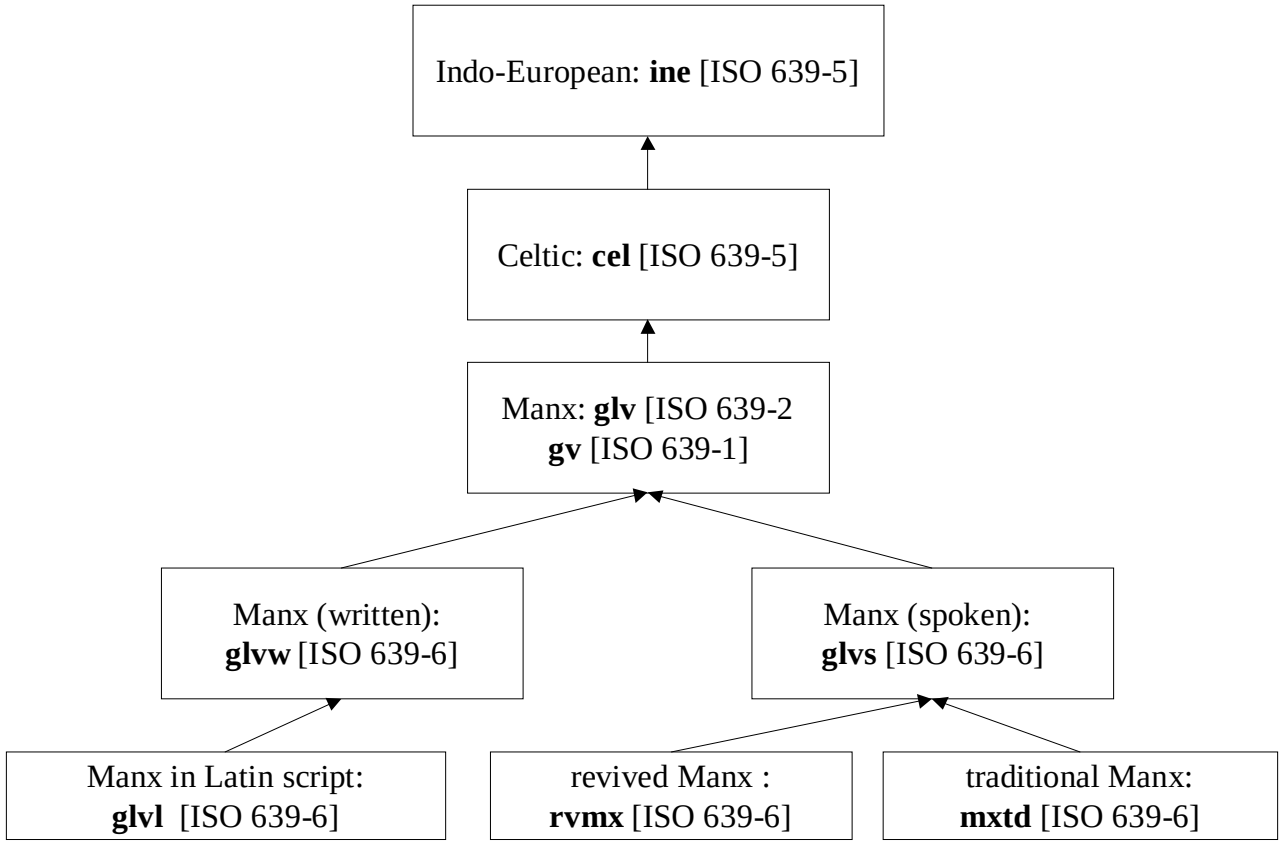
- Described in LDIF, and readily expressible using XML, where “identifier” is intended as a reference name for the language.
- “Reification”: we want both /identifier/ and the values of representations to be usable.

[**Data Category**]
[**Administration Identification**]
[**Administration Record**] /identifier/ = West Germanic
[**Description**] /broader data category = /Germanic languages/
[**Language Section**] /language/ = /en/
[**Name Section**] /name/ = West Germanic
[**Language Section**] /language/ = /fr/
[**Name Section**] /name/ = germanique occidental

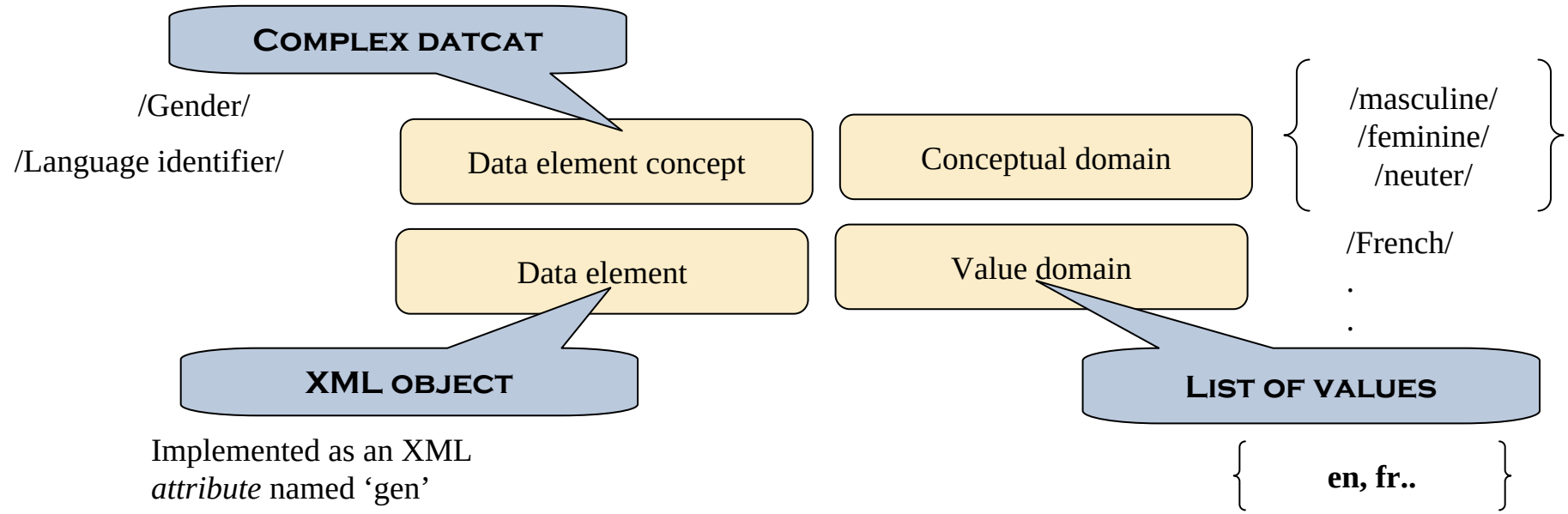
[**Data Category**]
[**Administration Identification**]
[**Administration Record**] /identifier/ = Germanic languages
[**Description**] /broader data category = /Indo-European languages/
[**Language Section**] /language/ = /en/
[**Name Section**] /name/ = Germanic languages
[**Language Section**] /language/ = /fr/
[**Name Section**] /name/ = germaniques, langues

[**Data Category**]
[**Administration Identification**]
[**Administration Record**] /identifier/ = Indo-European languages
[**Description**]
[**Language Section**] /language/ = /en/
[**Name Section**] /name/ = Indo-European languages
[**Language Section**] /language/ = /fr/
[**Name Section**] /name/ = indo-européennes, langues

The Language Documentation and Interchange Format (LDIF)



Integration



<w lemme=vert lang=fr gen=…>verte</w>

Integration

kca		Khanty (Khanti, Hanty, Xanty, Ostyak [Ethnologue / 639-3])
kcaw	kca	Khanty Written
kcal	kcaw	Khanty Written Latin Script
kcac	kcaw	Khanty Written Cyrillic Script
kcao	kcac	Khanty Written Cyrillic Script Obdorsk Model
kccc	kcac	Khanty Written Cyrillic Script Central Ob Model
kcav	kcac	Khanty Written Cyrillic Script Vakh-Vasyugan Model
kcak	kcac	Khanty Written Cyrillic Script Kazym Model
kcam	kcac	Khanty Written Cyrillic Script Vakh-Surgut-Shuryshkar

RFC 4545/6/7 and successors:

kcal => kca-Latn; kcac => kca-Cyrl

6kcam => [kcac = kca-Cyrl] => [kcaw] => kca.

Language Codes & Language Resources

- LIRICS partners are also delivering a portfolio of metamodels and metadata standards, at various stages of ratification by ISO, including:
  - ISO 24611 Morphosyntactic annotation framework (MAF)
  - ISO 24613 Lexical markup framework (LMF)
  - ISO 24615 Syntactic annotation framework (SynAF)
  - ISO 24617-1 Semantic annotation framework (SemAF) -- Part 1: Time and events
- These standards are complemented by test suites and an open-source implementation platform

http://lirics.loria.fr

Shapes of Things to Come

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*

Acknowledgements

- EU eContent project LIRICS (22236)
- ICT Marketing Ltd, Wales
- British Standards Institution
- UK's Science Research Infrastructure Fund (SRIF)
- Higher Education Innovation Fund (HEIF)
- Department for Trade and Industry's Knowledge Transfer Partnerships scheme (KTP 1739).
- Contributions and efforts of colleagues and peers in ISO, BSI, IETF, in the projects identified, and in the wider community also.