

# Developments in Language Codes standards

Lee Gillam<sup>1</sup>, Debbie Garside<sup>2</sup>, Chris Cox<sup>2</sup>

<sup>1</sup> Department of Computing, University of Surrey,  
Guildford, Surrey, UK

<sup>2</sup> GeoLang Ltd, Haverfordwest,  
Pembrokeshire, UK,

L.Gillam@surrey.ac.uk,  
{D.Garside, Chris.Cox}@GeoLang.com

**Abstract.** This paper discusses developments in international standards (ISO) for metadata that can be used to denote languages, the ISO 639 series of standards. These developments are intended to support improvements in the reusability and interoperability of data and provide for future generations of language resources and technologies. The focus of the paper is on aspects of the model for the Language Documentation and Interchange Framework (LDIF) for ISO 639 part 4. ISO 639 part 4 and the LDIF were initially intended for providing integration across ISO 639, with a view to supporting the future implementation of an integrated “standard as database” version of ISO 639.

**Keywords:** ISO 639; language identifiers; ISO standards; metadata; tags; multilinguality; interoperability; digital content; social networking; folksonomy

## 1 Introduction

The storage, management and retrieval of increasing amounts of data comprising spoken, written and signed languages of the world, places significant demands on precision in identification. Extant sets of language identifiers as encompassed in parts 1 and 2 of the ISO 639 standard for “Language Codes”, providing metadata identifiers for languages, can be used to denote around 400 of the world’s languages. The broad level of identification provided by these two standards has been suitable historically, but is increasingly considered to be of limited value for an emerging and diverse set of applications. The ISO 639 identifiers for languages are expected to expand from a set of around 400 to between 20,000 and 30,000, compiled from multiple encyclopaedic sources [1], [2], [3], [4], [5], [6]. This number, on a par with the Universal Decimal Classification (UDC) system, requires systematic interpretation and use, with relationships to other items specified.

In this paper we discuss some of the developments in the ISO 639 series of standards, particularly with respect to the model developed for the Language Documentation and Interchange Framework (LDIF) for ISO 639 part 4. ISO 639 part 4, currently in preparation, is intended to provide guidelines for implementation and management of the 5 currently conceived parts of ISO 639: the 2 current parts

identified above, and 3 subsequent parts at various emergent states. The framework provided was initially intended for integration of these parts of ISO 639 to ensure that the 639 series was systemic, and may be suitable for integration at some levels with other systems of language identifiers. Current initiatives within ISO to convert certain types of standards to database implementations places greater demands on ISO 639 part 4 and the LDIF to provide a specification suitable for an integrated database of language identifiers. We have discussed using LDIF to support an interchange format for language data elsewhere [7].

Gillam developed the LDIF under the auspices of the LIRICS<sup>1</sup> project in relation to a number of related initiatives, and LDIF is under international review in the current draft of ISO 639 part 4. LDIF is intended to be fully compatible and interoperable with emerging (meta-)models for lexical, terminological, syntactic, morphosyntactic and semantic annotation frameworks being developed under the auspices of ISO within the LIRICS project, and to provide a means for interlinking the collection of identifiers provided across the ISO 639 series.

## 2 Background

While scientists would like to be able to verify the findings of others using identical apparatus, in traditional laboratory environments and in *in silico* simulations and experiments, industrialists are interested in adding value to a variety of business content. Researchers at Berkeley have estimated that new data are produced at the annual rate of a couple of new *exabytes* (billion gigabytes). If what is produced in relation to laboratory work or in silico experiments or business content is left to reside in data tombs [8] of this scale, identified using bespoke and ad hoc means, the business or the scientific community will only be able to gain limited value from this and much of this potentially valuable data will eventually be rendered inaccessible.

Digital content is becoming both profuse and complex, and traditional boundaries between media types are eroding. Small-scale digital devices are now easily capable of supporting applications that integrate text with image and sound, enabling searching and retrieval of music, photographs and videos. User-generated content has occasionally been used in broadcast news, and is prevalent in the building of so-called social networks: next generation websites such as Flickr (still images), YouTube (video) and MySpace (individuals). Sites such as YouTube, referred to by some as Web 2.0, blur the boundary between web and television, blurred further by the availability of digital television over the internet (IPTV), and there has been discussion over whether legislation applicable to broadcasting should be applicable to these kinds of websites also. The various social networking initiatives have varying degrees of reliance on their contributors' use of metadata. Flickr, for example, provides a list of "Hot Tags", presented as a *tag cloud*, with a similar presentation of tags on a per-user basis. Those who have coincidentally selected the same tags may, or may not, find related images – there is no indication whether these tags are language specific, and semantics can only be implied by relation to other such tags.

---

<sup>1</sup> The EU e-Content project Linguistic Infrastructure for Interoperable Resources and Systems (LIRICS: eContent 22236). <http://lirics.loria.fr>

The term *folksonomy* has been coined for this kind of tag generation, which produces an open-ended and generally unstructured list of identifiers with a potentially long-tailed distribution of such tags.

Substantial efforts are still needed in the management of the exabytes of data being generated. The adoption of metadata standards, often limited to small sets such as the Dublin Core, may have been limited in part by the lack of availability and understanding of substantial enough collections of metadata identifiers, or metadata registries as identified by the ISO 11179 series of standards. Large-scale adoption of systems such as UDC may seem like overkill for small applications, although such an adoption could have substantial future potential. The use of small sets of metadata sometimes leads to “tag abuse”, where the limited set and desire of users to incorporate a greater amount of information come into conflict: a field intended for use as a date, in conformity with the international standard for date and time notation ISO 8601, could be filled with non-conformant values such as “last week”, requiring potential efforts in temporal analysis.

As the volume of digital content increases, for the addition of metadata to be beneficial beyond one or two users coinciding on tags by chance it needs to be done systematically, with the view that a systematic approach could lead towards automation. Furthermore, due consideration should be given to the preferences of users, in particular the language of the users, by supporting multilinguality and reducing ambiguity: in folksonomies, coincidental tags such as “cat”, shorn of context, could indicate a number of possibilities. Multiple users could be assisted in resource discovery by ensuring that they are adding *valuable* tags to their own resources, in identifying the language of the tag, and in documenting the intension of the tag. Lexicographers and terminologists, amongst others, would easily be able to propose solutions within such endeavours.

The long-term view that we have is of automatic indexing and metadata extraction for multilingual and multimedia content. With regard to multilingual content, the LIRICS project is developing a number of potentially beneficial international standards aimed at the reusability of language resources, with emerging (meta-)models for lexica, terminology, syntax, morphosyntax and semantic annotations. Work on expanding the ISO 639 “Language Codes” to cover increasingly granular and precise identification of languages is of interest in LIRICS for their use as basic descriptors for language resources. Automatic identification of language, and particularly linguistic diversity, would seem to be a vital initial element in organizing content within systems such as those proposed in LIRICS. Such systems must be capable of differentiating amongst written languages and, potentially, amongst the substantially challenging spoken form in, for example, audio channels and video content. Automatic indexing of ever-larger collections of multimedia content is under investigation in other research also being conducted at the University of Surrey<sup>2</sup>. A metadata registry of language identifiers [9] that helps in their management and use could also be of benefit for other scientific endeavours aimed at improving the

---

<sup>2</sup> This research is being undertaken in the UK Engineering and Physical Sciences Research Council (EPSRC) sponsored project on Recovering Evidence from Video by fusing Video Evidence Thesaurus and Video MetaData (REVEAL: GR/S98450/01) with the goal of automatic annotation of video captured from closed-circuit television cameras. <http://www.computing.surrey.ac.uk/ai/reveal>

technologies used for science and e-Science [10], particularly in the so-called Data Grids, or Knowledge Grids [11].

## 2.1 Current and Forthcoming Language Code Standards

The expansion of ISO 639 from a set of two standards encompassing around 400 identifiers, to a set of six standards encompassing over 30,000 identifiers is proceeding according to the schedule shown in Table 1. The timetable for publication is subject to change, and the drive for “standards as databases” may make further revisions to ISO 639 part 4 necessary for future-proofing purposes.

**Table 1.** Standards in the ISO 639 series on “Codes for the representation of languages” including dates of publication, organisation responsible for maintaining the integrity of the data, and the approximate number of identifiers, as currently conceived, included in or expected for the standard.

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
<i>ISO 639-3: Part 3: Alpha-3 code for comprehensive coverage of languages</i>	<i>Expected early 2007</i>	<i>Summer Institute of Linguistics (SIL)</i>	<i>7000</i>
<i>ISO 639-4: Part 4: Implementation guidelines and general principles for language coding</i>	<i>Expected late 2007.</i>	<i>n/a</i>	<i>n/a</i>
<i>ISO 639-5: Part 5: Alpha-3 code for language families and groups</i>	<i>Expected late 2007.</i>	<i>TBC</i>	<i>100</i>
<i>ISO 639-6: Part 6: Alpha-4 representation for comprehensive coverage of language variation</i>	<i>Expected early 2008.</i>	<i>GeoLang</i>	<i>25000</i>

## 2.2 Burden of Interpretation

ISO 639 parts 1 and 2, and latterly part 3, consist of lists of the identifiers and their names. The identifiers of these standards require systematic interpretation and use, yet relationships to other items within these lists are generally unspecified. The only implied relationships are between identifiers for historical varieties of languages, for example, separate identifiers for “German”, “Low German”, “Middle High German, (ca.1050-1500)” and “Old High German (ca.750-1050)”. These identifiers are presented discontinuously, as initially are those for part 3. By reference to the ISO 639-3 website<sup>3</sup>, it is possible to manually derive relationships from “German”

<sup>3</sup> ISO 639-3 data available at: <http://www.sil.org/iso639-3>

(identifier: *deu*) by using it as an entry point to the online version of the Ethnologue, in which it is referred to as “Standard German”. The entry for “Standard German” provides a set of human-readable information about the language and Ethnologue provides the language family tree in which “East Middle German” is the immediate class that also contains Lower Silesian (*sli*) and Upper Saxon (*sxu*). It is possible to explore these relationships further, however they are not immediately available for computational purposes in this form, and further information about temporal relationships appears equally unavailable.

ISO 639 parts 1 and 2 provide unstructured lists of identifiers, albeit relatively closed-ended; given the developments in progress, it is unlikely that either standard will undergo much in the way of expansion. The burden of interpretation and use sits squarely on the user: users can assume that they are able to infer what a given identifier stands for, though they should not. The oft-assumed mnemonic nature of some of these codes does not assist in avoiding such confusion, sometimes even amongst the experts.

Parts 5 and 6 of ISO 639 consist of hierarchically arranged identifiers that provide for some of the structure required to make associations between the identifiers of ISO 639, however work on the structure is currently in progress. The alpha-4 identifiers of ISO 639 part 6 are arranged to extend the identifiers of parts 1-3. ISO 639 part 5 provides a classification structure for the identifiers of parts 1-3, but to date the link between the classification structure and these identifiers has yet to be fully documented. When the links are available, it will be possible at minimum to generalize and specialize search queries over language resources that use any number of the ISO 639 identifiers.

The increased number of identifiers places demands on managing the identifiers. A catalogue of the names of all languages, classes and varieties could have somewhere upwards of 7000x7000 (49 million) entries, all of which have to be identified according to the language that they are in – the identifiers become self-consuming for such a system. Development of the supporting infrastructure for this multilingual catalogue is one of the public resources currently under discussion between the authors and the OmegaWiki project<sup>4</sup>.

### 3 LDIF Core and its Compatibility

The LDIF model was proposed and developed by LIRICS project participants and accepted for inclusion and review in the current version of ISO 639 part 4. The model was based on the need to be able to replicate the simplistic structure of ISO 639-1 and 639-2, to support the published documentation of languages in the Ethnologue, and to support the model being developed by the authors of this paper through BSI for ISO 639-6, adapted generalized and cross-validated from encyclopædic sources [1-6]. The intention was that the various sets of identifiers for languages (2-letter, 3-letter and 4-letter) provided by the ISO 639 standards would be compatible, interoperable,

---

<sup>4</sup> OmegaWiki aims to provide a publicly usable infrastructure supporting language resources, including information about languages, terminologies, dictionaries, and so on. <http://www.omegawiki.org>

mutually comprehensible, usable to varying degrees of precision, and support simple query expansion. ISO 639 has been developed, and continues to be developed, through ISO technical committee 37 (TC 37). The first standard to emerge from TC 37 that specified a metamodel was ISO 16642 for a Terminological Markup Framework (TMF). TMF emerged largely from the SALT project<sup>5</sup> and provided a mechanism for specifying interchange formats by integration with a set of metadata identifiers for terminology, referred to as Data Categories [12]. These Data Categories had been documented some years previously as ISO 12620; data categories may be referred to also as administered items, in accordance with ISO 11179. The approach taken in TMF has been adopted in large part within the further metamodels for annotation of language resources, and is evident in LDIF also.

The model for LDIF provides an expansion of, and compatibility with, the model in the latest revision of ISO 12620, and as such is expressible using XML. The ISO 12620 model incorporates metadata for describing metadata. The language identifiers of ISO 639 can be described using many of these metadata in specific ways. Key components of the ISO 12620 model, denoted in italics, and the use to which the ISO 12620 metadata descriptors can be used for describing language identifiers include the following:

- **identifier**: descriptor used for the unique identification of the data category (administered item) in a given context. This essential component is part of the *Administration Record*, with the context provided by information about **registration authority** and the **version**.
- **name**: a language-dependent denotation for the data category. Part of the *Name Section* of the data category. The *Name Section* can be used to further document the name, and to provide a multilingual thesaurus of language names, including those whose use may be deprecated for a variety of reasons. Since **name** is language dependent, information about the languages in which the name is appropriate should be documented also; this documentation in itself will demonstrate and necessitate the use of language identifiers.
- **broader concept generic**: an association made from a data category to the identifier of a more general data category. This optional component is part of the *Description*. *Description* also incorporates the *Name Section*.

The descriptors for **identifier** and **name** appear sufficient to represent much of what is currently available for ISO 639 parts 1 and 2 – particularly the names of languages in both English and French. The **broader concept generic** allows for hierarchical arrangement of identifiers, supporting simple structural needs of parts 5 and 6. However, since **identifier** has the condition of being unique in a given context, an immediate issue would arise in using the alpha-3 as the identifier in ISO 639-2 since some elements have dual alpha-3 forms for historical reasons. Furthermore, there are a number of ISO 639-2 alpha-3s that have equivalent ISO 639-1 alpha-2s. This is not problematic for a given **registration authority**, however the equivalence relationship should be registered somewhere for purposes of interoperability. In ISO 639-3, **identifier** has been referred to as **reference name**, a

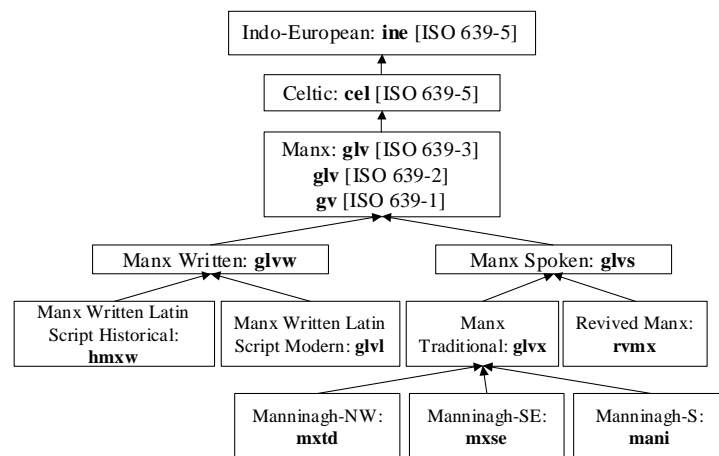
---

<sup>5</sup> The EU 5<sup>th</sup> Framework project Standards-based Access to Lexicographical & Terminological multilingual resources (IST-1999-10951): <http://www.loria.fr/projets/SALT/>.

selected name of the language; no preference for this name over any other is intended. The ISO initiative on standards as databases suggests that the notion of a registration authority should preferably exist at the level of the ISO 639 rather than at the level of each registering authority. This suggests the following requirements:

- **representation:** a descriptor for documenting the variety of equivalent identifiers for essentially the same item across, at least, ISO 639 parts 1, 2 and 3. A separate *Representation Section* allows for each representation to be further described. This allows for alpha-2, alpha-3, alpha-4, numeric, alphanumeric, hexadecimal, and any other constructed representational formats, providing for multiple language-independent forms not covered by **name** and not specific to any given application.
- **identifier:** this descriptor should be unique *across ISO 639*, as opposed to being unique within one of the registration authorities, such that **broader concept generic** can provide associations to more general data categories across the range of data available in the series of standards. Ensuring this consistency in identification is a work-in-progress across ISO 639.

An example is provided for Manx (Fig. 1), with multiple alpha-2 and alpha-3 identifiers, and associations made to language families of the current draft of ISO 639-5, and varieties of Manx according to ISO 639-6 (4-letter identifiers), that can be supported through such constructions.



**Fig. 1.** Integration of data from ISO 639 for “Manx” including the alpha-4 identifiers of ISO 639 part 6. The data for ISO 639-6 is currently undergoing verification and validation by GeoLang, so names and identifiers may be subject to modification.

Beyond names, identifiers and representations, further documentary information exists for each language. Traversing the online Ethnologue one finds that a variety of other metadata identifiers could be used to aid discovery, not limited to information about countries in which the language was or is spoken for which the ISO 3166 country codes could be employed. LDIF specifies a *Documentation Section* that could be used to provide for information related to geography, society, culture,

religion and so forth. A number of related metadata descriptors have been included in relation to LDIF in ISO 639-4 that may be useful for this purpose. A fully machine-readable representation for all such information, that would support query expansion in applications of automatic annotation, is a longer-term goal.

## **4 Discussion and Future Work**

While language experts may identify their content in a highly precise manner that helps in its retrieval, non-experts creating folksonomies may not know such specifics but could still be interested in using such identifiers to help others find their content better; the more precise the identification, the better the chances of more-easily accessing the desired content. Efforts on standardisation of metadata aim towards such a provision. Such efforts potentially enable convergence of markup formats also, improving the possibility of data reuse by allowing something akin to transcoding. One of the successes of XML has been the ease with which it can be used to produce interchange formats, however the resulting profusion of interchange formats necessitates meta-level interchange formats, and the various meta-models are intended to enable this.

LDIF is one means to integrate language identifiers defined by the ISO 639 series of standards. Such a model is important both for ISO 639 itself, but more so since these data are used indirectly by efforts such as XML: the current XML specification references the Internet Engineering Task Force (IETF) Request for Comments (RFC 3066). RFC 3066 combines the language codes of ISO 639 parts 1 and 2 with country codes of ISO 3166 to support specific semantics. This provides relatively familiar identifiers such as “en-US”, somehow denoting “American English”. RFC 3066’s successor RFC documents (RFC 4545, 4646 and 4547) expand this notion by inclusion of identifiers for language scripts from ISO 15924; work is in progress on further successor documents to incorporate the impending publication of ISO 639 part 3. Removal of script and geography extensions can be used to enable some broader information retrieval tasks, and some further aspects of tag matching are also documented. The approach is generative, however: a particular identifier composed of language identifier from ISO 639 part 1, 2 or 3, a country code from ISO 3166 part 1 and a script identifier from ISO 15924 may or may not result in a valid combination. ISO 639 part 6 data could be used to validate the combination, or, for example, to situate the combination with respect to alternatives, providing a variety of entry points to the ISO 639 system. Eventually, one presumes, XML and the wider user community will be able to benefit from all of these efforts and also to begin generating content that makes use of them. The authors of this paper are collaborating with the OmegaWiki project to provide portals that enable user communities to generate content that demonstrates the linguistic and cultural wealth of these communities, providing a point of presence for all of the languages of the world; this content will provide further documentary information for languages and their varieties, providing for a self-documenting system of identification.

**Acknowledgments.** This work has been supported, in part, by seed-funding provided by ICT Marketing Ltd, Wales and the British Standards Institution, by the EU eContent project LIRICS (22236), the UK's Science Research Infrastructure Fund (SRIF), Higher Education Innovation Fund (HEIF), and the Department for Trade and Industry's Knowledge Transfer Partnerships scheme (KTP 1739). The authors have contributed significantly to the ISO 639 community in recent years with respect to the emerging standards via BSI and ISO, and acknowledge the contributions and efforts of colleagues and peers in ISO, BSI, IETF, in the projects identified, and in the wider community also.

## References

1. Gordon Jr, R. G (ed.): *Ethnologue: Languages of the World*, 15th Edn. SIL International. (2005).
2. Dalby, D.: *Linguasphere Register of the world's languages and speech communities*. Linguasphere Press. (1999).
3. Voegelin, C.F. and F.M.: *Classification and index of the world's languages*. New York, NY: Elsevier North Holland, Inc. (1977).
4. Ruhlen, M.: *A guide to the world's languages*. Vol.1: Classification. London: Edward Arnold. (1987)
5. Bernard Comrie (ed.): *The World's major languages*. Oxford University Press, New York. (1987).
6. Chambers, J.K., Trudgill, P.: *Dialectology*. Cambridge: Cambridge University Press. (1998).
7. Gillam, L., Garside, D., Cox, C.: Information volumes and linguistic diversity: meeting the challenges for content management. In *Proceedings of 3rd International Conference on Terminology, Standardization and Technology Transfer (TSTT)*, 25-26 August, Beijing, PRC. (2006).
8. Fayyad, U., Uthurusamy, R.: Evolving data mining into solutions for insights. *Communications of the ACM* 45(8):28-31. (2002)
9. Gillam, L.: Metadata descriptors: ISO standards for terminology and other language resources. *Proceedings of the 1st International e-Social Science Conference*. (2005).
10. Coveney, PV (ed.): *Scientific Grid Computing*. *Philosophical Transactions of the Royal Society of London Series A*, Volume 363. (2005)
11. Cannataro, M. and Talia, D.: The knowledge grid. *Communications of the ACM* 46 (1), 89-93. (2003).
12. Gillam, L., Ahmad, K., Dalby, D., Cox, C.: Knowledge Exchange and Terminology Interchange: The role of standards. In *Proceedings of Translating and the Computer* 24. (2002).