

A metadata infrastructure using ISO standards

Lee Gillam

Department of Computing, University of Surrey, UK

Abstract

This paper presents an overview of a number of international standards for metadata descriptors that have been, or are being, defined through ISO. ISO standards can be used for the creation of so-called metadata registries that encompass both the metadata databases (registers) and the management policies and processes that govern their use. The standards for metadata registries are being used as a basis for infrastructure for metadata storage and retrieval for the Language Resources community as part of an EU eContent project. This paper describes metadata registries and their use for supporting the documentation of extant metadata and the specification of new metadata and interoperable interchange formats.

1. Introduction

Scientists of various hues are at different stages of discovering the potential benefits of large-scale e-mediated scientific experimentation that makes use of distributed and heterogeneous computer systems, sometimes across administrative domains. The ability of scientists to prove, disprove or produce results comparable to others relies on the access of the scientists to similar, or preferably identical, tools, techniques, and data. Comparability and repeatability are cornerstones of scientific work. In typical scientific writing, scientists fix on a single dataset and one or more techniques, or fix on a technique and discuss its benefits using one or more datasets, hoping to prove something in the data or something about their preferred technique. Use and reuse of datasets, in general and perhaps in Grid initiatives, relies to some extent on the ability of the user, or the user's program, to process the dataset, and by extension to have some knowledge of the structure of the data in the dataset. Though spreadsheet-type formats are not uncommon (data in comma-separated or otherwise delimited formats), data are increasingly produced for wider distribution that are described, or annotated, using the eXtensible Markup Language (XML: Bray et al 2000). The markup of these data can be in-line, where the markup is interspersed with the data, or in a stand-off configuration where, often, character offsets of the original data are used but the original data is left as is. One of the benefits of XML has been the ease with which it can be used to produce interchange formats, at the cost of interchange format profusion. A particular XML format for data storage or data interchange is constructed, or authored, on the

basis of "understood" identifiers that may be restricted by datatypes or data ranges identified in schemas. Disparate formats, defined by groups undertaking research in distributed laboratories, may use these identifiers in compatible ways; use of standardized metadata, where the descriptions are consensually agreed upon, enables the convergence of mark-up formats and improves the reusability of data.

This paper presents an overview of international standards, either already in existence or currently emerging through ISO, for metadata. These standards include those for so-called metadata registries that enable the documentation and management of metadata. Metadata registries, defined in the six parts of ISO 11179, potentially enable the convergence of markup formats and improve data reusability. Here, we consider a need for a descriptive approach, documenting that which already exists, as a step prior to be able to prescribe a standard format.

Examples are provided of how these standards are being used by communities dealing with documentation of information about languages. In particular, how these will be used as a basis for better documentation of the ISO standards series on "Language Codes", ISO 639. ISO 639 is a commonly, but sometimes haphazardly, used set of identifiers that is currently undergoing expansion from 2 parts to 6 parts - from about 400 identifiers to well over 30,000. Metadata registries are discussed with reference to a number of other international standards that make use of them to provide for mark-up and interchange, including those for domain-specific terminology collections which may be useful elsewhere as a basis for ontologies (Gillam, Tariq and Ahmad 2007).

2. Grids, metadata and markup

Foster and Kesselman identified standards as being important for provenance, quality and validity, “correct” combinations of metadata, alternate representations, personalisation and interchange (1999. 105-6, 123-7). Within Grid initiatives, uptake of ISO standards often appears to be either well-hidden or ignored, perhaps due to inertia, unfamiliarity, or perception that updates to ISO standards cannot keep pace with other standards being produced for the technology. Arguably, the periods of stability offered by ISO standards is essential for even greater uptake of Grid technologies, for example within the business community.

The cost of purchasing (a number of cross-referring) ISO standards, and the perceived impenetrability of the contents of these documents, may be one of the barriers to widespread ISO standards adoption. To ensure conformity and/or compliance with a number of standards for basic identifiers, one would need to digest, and in some cases purchase, standards for quantities (ISO 31-0), language codes (ISO 639-1, ISO 639-2), country codes (ISO 3166-1), codifications of dates (ISO 8601), character sets (ISO 10646-1), and data types (ISO 11404). Many of these standards are incorporated in others, for example those of the Internet Engineering Task Force (IETF): the combination of ISO 639-1, ISO 639-2 and ISO 3166-1 existed for some time in the IETF’s Best Common Practice (BCP) 47 (RFC 3066 to September 2006). In turn, XML uses BCP 47 as its language tag syntax (xml:lang). However, good understanding of the ISOs is still important to ensure good use of this syntax.

Metadata such as in the Dublin Core are often used, but insufficient granularity can result in what some refer to as “tag abuse”. Use of a “date” field is open to some interpretation, unless insistence on the field being filled with data in a format compatible with ISO 8601 – a “date” field filled with “yesterday” requires greater effort for the machine interpretability expected of XML. Furthermore, the tendency to produce language-based identifiers can result in, for example and at minimum, French-based identifiers for author (auteur) and gender (genre), with potential for some incorrect assumptions for the latter if taken out of context; a “conceptual” identifier, or namespace, is required to ensure that differences in denotation, or naming, are easily handled.

ISO metadata standards, in particular a metadata registry, could be used to underpin

Grid initiatives, with particular thought given to program composition in Semantic Grids (de Roure, Jennings and Shadbolt 2003) and data mining-type applications in Knowledge Grids (Cannataro and Talia 2004). At the interface between programs, potentially provided by research groups worldwide, reference to common metadata sets would need to be made, and most likely enforced. On the basis of the extant ISO metadata standards, it is possible to consider an e-Science metadata registry that standardizes and centralizes the problem and provides best practice for the community: one-to-one interchange can be effective, but many-to-many suggests the need for such an infrastructure.

3. Overview of Metadata standards and Language Resources

3.1 ISO 11179 for metadata registries

The ISO 11179 series of standards define a metadata registry as “a database of metadata that supports the functionality of registration”, where registration covers the rules, operations, and procedures. Each item in a metadata registry is uniquely identified, and documented for provenance. The unique identification is formed of three components: (i) an identifier for the registry; (ii) an identifier for the metadata; (iii) a version identifier for the metadata. Hence, identical identifiers and versions can remain separately defined across different registries. A specific “use” of the metadata, then, needs to document the unique identifier of which it is an instance so that interoperability can be assured. Dublin Core metadata items are prefixed “dc.” as an attempt to enforce this. XML namespaces are used to provide common reference to schemas, but this only results in placing the standardization issue elsewhere. The XML attribute “xml:lang” is a particular case in point of this: xml:lang currently¹ relies on interpretation of other standards for its values, and potentially a number of assumptions have to be made about the *intended* meaning of the contained identifiers.

ISO 11179 parts 1 to 6, identify the promotion of the following:

¹ XML 1.0 (Fourth Edition), 29 September 2006, currently refers to IETF RFC 3066 rather than BCP 47; RFC 3066 uses ISO 639 and 3166, while BCP 47 also incorporates identifiers from ISO 15924. To “understand” xml:lang, it is necessary to work through these underlying standards.

- Standard description of data: this allows for searchability of the metadata to reduce the likelihood of the registration of multiple identifiers for the same purpose
- Common understanding of data across organizational elements and between organizations: the description assists in the use of the metadata in the same way that business terminology should be managed
- Re-use and standardization of data over time, space, and applications: at minimum, different versions of the metadata enable a view and history of the use of the metadata.
- Harmonization and standardization of data within an organization and across organizations: different datasets that converge on metadata items begin to become used in combination.
- Management of the components of data
- Re-use of the components of data

The six parts of the ISO 11179 series handle different aspects of the management of metadata, with metadata being:

- Classified according to part 2
- Specified according to part 3
- Defined according to part 4
- Named according to part 5
- Registered according to part 6.

All of the processes outlined provide particular benefits in understanding how the metadata should be managed and used. For example, the “acceptability” of a particular item of metadata can be denoted by its registration status according to part 6, by which it can be defined as “standard”, or “deprecated”.

ISO 11179 documents a separation between the “concept” and the “use”. Figure 1, below, is taken from part 4 and shows the relationship between an abstract notion (Data Element Concept), the things that itemize the abstract notion (Conceptual Domain), the representation of the abstract notion (Data Element) and the values for the element (Value Domain).

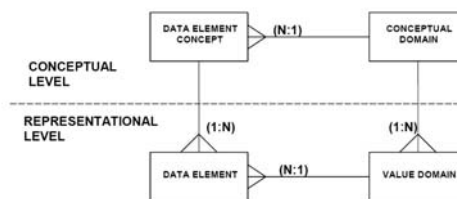


Figure 1: Relationship between abstract notion of a metadata item and the values that exist for it

By way of example, we can consider the existing list of identifiers provided for by ISO 639 parts 1 and 2, containing codes for (the representation of names of) languages, so the conceptual domain is, in essence, the list of the “languages of the world”:

ISO 639 parts 1 and 2 can be considered to have, at least, four value domains as shown below in Table 1: the alpha-3 (3-letter) identifier, the alpha-2 (2-letter) identifier and the English and French names for these.

Alpha-3	Alpha-2	English name	French name
eng	en	English	anglais
epo	eo	Esperanto	espéranto
est	et	Estonian	estonien
ger / deu	de	German	allemand
gla	gd	Gaelic; Scottish Gaelic	gaélique; gaélique écossais
jpr		Judeo-Persian	judéo-persan

Table 1: Sample data from ISO 639 parts 1 and 2

It is possible to consider, therefore, that applications could make various use of these 4 different Value Domains. The picture is, however slightly more complex than at first appears: there are less values in the value domain for alpha-2 than for alpha-3: there is no equivalent alpha-2 value for “jpr”. In some instances, there are 2 alpha-3 identifiers for the languages due to historic reasons. Furthermore, there can be one or more names for the language in each language. The underlying model for these brief lists of identifiers is, therefore, a little more complex than it appears, so making use of even the alpha-2 and alpha-3 identifiers, and presenting human-readable descriptions to end users, requires a little more thought.

A similar example is given in the ISO 11179 standards for the 7 value domains for countries of the world from ISO 3166.

3.2 Language Resources

The term “language resource” is applied to a variety of data and programs, from glossaries to ontologies, from natural language processing (NLP) applications and data to multimedia databases, and applying to spoken, written and signed forms of communication, and the software that processes them.

The Language Resources community has a history of working on metadata initiatives including, but not limited to, initiatives such as IMDI (Wittenburg, Broeder and Sloman, 2000), American Open Language Archives Community (OLAC) (Bird and Simons 2000), the Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard 1999) and the XML Corpus Encoding Standard (XCES) (Ide, Bonhomme and Romary 2000).

More recently, standardization efforts for Language Resources have appeared under the ISO banner, with developmental efforts receiving some funding under the EU’s eContent programme. These initiatives are emerging through ISO technical committee 37 (TC 37). The Linguistic Infrastructure for Interoperable Resources and Systems (LIRICS) project is providing for the development of standards for lexica, syntax, morphosyntax and semantic annotations aimed at reusability of these kinds of language resources. This work is likely to be taken up, in part, if funding is secured by a consortium for an EU research infrastructure for a Common Language Resources and Technology Infrastructure (CLARIN) intended to offer services to linguistic, humanities and social science communities.

The approach being taken for the lexical, syntax, morphosyntax and semantic resources builds on that for the production of terminology markup. This approach is outlined in the “Terminological markup framework” (TMF, ISO 16642) (Gillam et al 2002), the first standard to emerge from TC 37 that specified a metamodel. ISO 16642 specifies a metamodel that abstracts away from specific implementations and reflects thoughts of terminologists that a “concept” (Terminological Entry) can be expressed in 1 or more languages (Language Section) by 1 or more terms (Term Section), as shown in Figure 1 below, from ISO 16642.



Figure 2: ISO 16642 Terminological Metamodel

The association of metadata identifiers with an instantiation of the metamodel can be used either to document existing formats, whereby the interoperability with related formats can be determined, or to ensure that a newly constructed format is compatible. TMF emerged largely from the EU-sponsored Standards-based Access to Lexicographical & Terminological multilingual resources (SALT) project. The metadata elements themselves are described in ISO 12620, although this standard is in the process of revision as outlined in the next section. By reference to **language identifier**, **term** and **definition** in ISO 12620, a minimal terminology model would associate a **language identifier** with the Language Section, a **term** with the Term Section, and a **definition** with either the Terminological Entry (concept definition) or the Language Section (definition for the language), with further metadata as required for documentation purposes. The metadata identifiers would have specific instantiations in an XML outline, for example as “lang”, “def” and “term”.

A simple, compatible XML outline would look something like:

```

<TE> <def></def> <LS lang="">
<TS> <term></term> </TS> </LS>
</TE>
  
```

Collections of terminology, so-defined, can be used as the basis for ontologies. Interchange can be mediated at the level of the standard, and the outline structure of such an ontology can be learnt automatically from collections of written texts (Gillam, Tariq and Ahmad 2007, Gillam 2004).

3.3 Metadata for Language Resources: ISO 12620

ISO 12620, as published in 1999, contains around 170 metadata identifiers, referred to as Data Categories, for the management of computer collections of multilingual terminology. The revision of ISO 12620 is currently underway, aiming at the provision of a Metadata Registry for Language Resources in

which these identifiers will be housed². ISO 12620 specifies a model for metadata based on ISO 11179, that has some similarities with that for terminology in Figure 1. These Data Categories can be considered as Administered Items in accordance with 11179. The Data Category Interchange Format (DCIF) is also specified for interoperability with other metadata registries. The ISO 12620 model allows for the documentation of identifiers used in language-specific ways (Language Section), and the documentation of multiple names for these metadata (Name Section [Term Section]) in language-specific ways. Furthermore, certain languages place specific requirements on Conceptual Domains and the Value Domain, for example, the notion of “grammatical gender” may take values for “masculine”, “feminine” and “neuter” in French and German, other languages may have more or less, but the notion itself is essentially missing from English.

3.4 Metadata for Language Documentation: ISO 639

ISO 639 has also been developed, and continues to be developed, through the same technical committee (TC 37). ISO 639 parts 1 and 2 have, for many years, provided an unstructured list for around 400 languages. A recent part 3, published in early 2007, provides for substantially greater coverage, bringing the number to upwards of 7500. The data are again provided in an unstructured fashion, though human interpretation is possible by reference to, and traversal of, the source of the data. Still, much of the burden of interpretation and use is placed on the user.

Further developments are underway in providing ISO 639 parts 5 and 6, comprising hierarchically arranged identifiers that provide the structure needed to make machine-processable associations between, and interpretations of, the identifiers of ISO 639. ISO 639 part 6 specifies alpha-4 identifiers, arranged to extend the identifiers of parts 1-3, and brings the total number of metadata identifiers to over 30,000. ISO 639 part 5 provides a classification structure for the identifiers of parts 1-3, but the link between the classification structure and these identifiers remains to be fully documented. When the full series is available, expected some time in 2008, it will be possible at minimum to generalize and specialize search queries over language resources that use these identifiers.

The increased number of identifiers places further demands on their management. Consider, for example, that a simple catalogue of the names of all languages in parts 1-3 has potential for, at least, 7500x 7500 entries (> 56 million). Consider, also, that for adequate documentation purposes, the system needs to be self-documenting such that we know the language of each name. Development of the supporting infrastructure for this multilingual catalogue is one of the public resources currently under discussion with the OmegaWiki project³.

The management of identifiers of ISO 639, and especially the supporting documentation about the languages of the world, is the responsibility of ISO 639 part 4. This standard provides an expansion of the ISO 12620 model and a Language Documentation (and Interchange) Format (LDIF) that is compatible with the ISO 12620 DCIF and expressible using XML. The LDIF model emerged from the LIRICS project and is based on the need to be able to replicate the “simplistic” structure of ISO 639-1 and 639-2, and to support the further parts of the ISO 639 series. The purpose of the model is to ensure that the different identifiers for languages (2-letter, 3-letter and 4-letter) provided by the full set of ISO 639 standards can be used in compatible, interoperable and mutually comprehensible ways.

Development of the metadata registry for the languages of the world is very much a work in progress, some details of which have been presented elsewhere (Gillam, Garside and Cox 2006, 2007). The move from 400 identifiers to over 30,000 necessitates such a provision, and this is given further impetus by moves within ISO towards the provision of “standards as databases”. As such, ISO 639 part 4 also provides the model for an ISO 639 database standard.

The processes of verification, validation and continued maintenance of such a large collection of metadata requires a community effort. This is being effected as follows: the British Standards Institution (BSI) and GeoLang have responsibility for ISO 639 part 6, the “paper” standard, and the initial provision of the supporting conformant data and registry; an international expert panel has been formed, the World Language Documentation Centre (WLDC), as arbiters of quality; and community contributions and debate will be managed using OmegaWiki. OmegaWiki will be providing portals per identifier in which data about the

² A prototype of this Metadata Registry is currently offered at: <http://syntax.inist.fr>

³ <http://www.omegawiki.org>

languages, such as basic vocabularies, can be collated. Learning from some of the difficulties of Wikipedia, the intention is that experts of the WLDC will ensure that the contents of these portals are treated respectfully. The need for further identifiers, or modifications to the existing set is expected to emerge through these portals, and GeoLang will undertake to ensure inclusion does not impact significantly on stability over time of the entire system. This is, indeed, an effort that presents substantial challenges, and promises much debate.

4. Discussion

Metadata Registries are being developed, via ISO, to support a variety of research centred around Language Resources. These registries contain metadata for the documentation of languages themselves, as well as for the documentation of analysis of the languages, including syntax, morphosyntax, lexical, terminology and semantic information. These registries are based on the existing ISO series of standards for metadata registries, ISO 11179. Two compatible registries are expected, one entailing the documentation of languages in general, one for components of language. The documentation of languages provides for multiple possible value domains, however for use in documentation of language components, a single Value Domain is expected to be used.

The expansion of the ISO 639 “Language Codes” series of standards will cover increasingly granular and precise identification of languages. The multilingual thesaurus of languages names, suggested above, requires some precision in identification that may be greater than that suggested: the use of multiple scripts for certain languages would require that the language and script for the name be documented – and, even further, that this name is used in certain countries but not others. The full set of metadata needed simply to document the names of languages could be quite substantial before one even gets to the documentation of the language itself.

Automatic identification of language, and particularly linguistic diversity, is a longer term, and necessary, goal: the 400 identifiers of ISO 639 parts 1 and 2 are currently used with limited accuracy. Knowing accurately which of 30,000 identifiers is appropriate is not generally within human capabilities. The research community has developed a variety of systems for identifying the languages of the web, and for some speech differentiation. With digital audio

and video content so prevalent, such applications become increasingly useful.

At minimum, a metadata registry of language identifiers could also be of benefit for other scientific endeavours aimed at improving the technologies used for science and e-Science.

Defining and using ISO standard-conformant metadata has the potential, longer-term, for improving management of data and facilitating the scaling up of research. Data described using finer-grained metadata, and metadata for which more general or more specific resources can be discovered may be useful beyond their original composition, reducing the possibility of even more Data Tombs (Fayyad and Uthurusamy 2002).

An e-Science community metadata initiative at a national or international level could, amongst other things:

- Provide increased ease of use of extant data sets.
- Ensure common understanding of data across research groups and, especially, in virtual organizations
- Provide for data reusability across applications
- Enable convergence of different datasets that would otherwise be isolated.
- Contribute to overcoming a “major gap” previously identified in e-Science at large (e-Science Gap Analysis, 2003, p10),

A resulting e-Science Metadata Registry would provide a baseline for, at minimum, Data Grids – and integration with a virtualization system such as the Storage Resource Broker (SRB) via its Metadata Catalog (MCAT) would provide a useful demonstration of the benefits of such a Registry.

Acknowledgments. This work has been supported, in part, by the EU eContent project LIRICS (22236), and the Department for Trade and Industry’s Knowledge Transfer Partnerships scheme (KTP 1739). The author has contributed significantly through ISO, and in particular to a number of the standards mentioned, and acknowledges the contributions and efforts of colleagues and peers in ISO, BSI, IETF, in the projects and initiatives identified, and in the wider community also. The author is grateful to the three anonymous reviewers for their comments.

References

- Bird, S. and Simons, G. (2000) "White Paper on Establishing an Infrastructure for Open Language Archiving". <http://www.language-archives.org/docs/white-paper.html> (20 Apr. 07)
- Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E. (eds.), (2000). "Extensible Markup Language (XML) Version 2.0". W3C Recommendation. <http://www.w3.org/TR/REC-xml>
- Cannataro, M. and Talia, D. (2004). "Semantics and Knowledge Grids: Building the Next-Generation Grid". *IEEE Int.Sys* 19(1), pp56-63
- de Roure, D., Jennings, N. R. and Shadbolt, N. (2003) "The Semantic Grid: A future e-Science infrastructure". In Berman, F., Fox, G. and Hey, A. J. G., (Eds.) *Grid Computing - Making the Global Infrastructure a Reality*. pp. 437-470. John Wiley and Sons Ltd.
- Fayyad, U. and Uthurusamy, R. (2002) "Evolving data mining into solutions for insights". *Communications of the ACM* 45(8), pp28-31
- Foster, I. and Kesselman, C. (Eds.) (1999) "The Grid: Blueprint for a New Computing Infrastructure". Morgan-Kaufmann: San Francisco, California.
- Fox, G. and Walker, D. (2003) "e-Science Gap Analysis" <http://grids.ucs.indiana.edu/ptliupages/publications/GapAnalysis30June03v2.pdf>. (20 Apr. 07)
- Gillam, L., Ahmad, K., Dalby, D. and Cox, C. (2002) "Knowledge Exchange and Terminology Interchange: The role of standards". *Proc. of Translating and the Computer* 24. ISBN 0 85142 476 7
- Gillam, L. (2004). "Systems of concepts and their extraction from text". Unpublished PhD thesis, University of Surrey.
- Gillam, L., Garside, D., Cox, C.: (2006) "Information volumes and linguistic diversity: meeting the challenges for content management". In *Proc. of 3rd Intl. Conf. on Terminology, Standardization and Technology Transfer (TSTT)*, 25-26 August, Beijing, PRC.
- Gillam, L., Garside, D. and Cox, C. (2007) "Developments in Language Codes standards". In Rehm, Witt and Lemnitzer (eds.): *Data Structures for Linguistic Resources and Applications*. *Proc. of GLDV 2007*, 11-13 April 2007, Tübingen, Germany, Tübingen: Gunter Narr Verlag. ISBN 978-3-8233-6314-9
- Gillam, L., Tariq, M. and Ahmad, K. (2007). "Terminology and the construction of ontology". In *Application-Driven Terminology Engineering*, Ibekwe-SanJuan, F., Condamines, A. and Cabré Castellví, M-T. (eds.), 49-73.
- Ide, N., Bonhomme, P. and Romary, L. (2000) "XCES: An XML-based Standard for Linguistic Corpora". *Proc. of the Second Language Resources and Evaluation Conference (LREC)*, 825-30.
- ISO 31-0 (1992) "Quantities and units — Part 0: General principles". ISO, Switzerland.
- ISO 639-1 (2002) "Codes for the representation of languages Part 1: Alpha-2 code"
- ISO 639-2 (1998) "Codes for the representation of languages Part 2: Alpha-3 code"
- ISO 639-3 (2007) "Codes for the representation of languages Part 3: Alpha-3 code for comprehensive coverage of languages", ISO, Switzerland
- ISO 639-4: "Codes for the representation of languages Part 4: Implementation guidelines and general principles for language coding". ISO, Switzerland (forthcoming)
- ISO 639-5: "Codes for the representation of languages Part 5: Alpha-3 code for language families and groups". ISO, Switzerland (forthcoming)
- ISO 639-6: "Codes for the representation of languages Part 6: Alpha-4 representation for comprehensive coverage of language variation". ISO, Switzerland (forthcoming)
- ISO 3166-1 (2006) "Codes for the representation of names of countries and their subdivisions -- Part 1: Country codes". ISO, Switzerland.
- ISO 8601 (2004) "Data elements and interchange formats -- Information interchange - Representation of dates and times". ISO, Switzerland.
- ISO/IEC 10646-1 (2000) "Information technology -- Universal Multiple-Octet Coded Character Set (UCS) -- Part 1: Architecture and Basic Multilingual Plane". ISO, Switzerland.
- ISO/IEC 11404 (1996) "Information technology -- Programming languages, their environments and system software interfaces -- Language-independent datatypes". ISO, Switzerland.
- ISO/IEC 11179-1 (2004) "Metadata registries (MDR) - Part 1: Framework". ISO, Switzerland.
- ISO/IEC 11179-2 (2005) "Metadata registries (MDR) - Part 2: Classification". ISO, Switzerland.
- ISO/IEC 11179-3 (2003) "Metadata registries (MDR) - Part 3: Registry metamodel and basic attributes". ISO, Switzerland.

ISO/IEC 11179-4 (2004) “Metadata registries (MDR) - Part 4: Formulation of data definitions”. ISO, Switzerland.

ISO/IEC 11179-5 (2005) Metadata registries (MDR) - Part 5: Naming and identification principles”. ISO, Switzerland.

ISO/IEC 11179-6 (2005) Metadata registries (MDR) - Part 6: Registration”. ISO, Switzerland.

ISO 12620 (1999) “Computer Applications in Terminology – Data categories”. ISO, Switzerland.

ISO 16642 (2003) “Computer Applications in Terminology – Terminological markup framework (TMF)”. ISO, Switzerland.

Sperberg-McQueen, C.M. and Burnard, L. (eds.). (1999). “Guidelines for Electronic Text Encoding and Interchange. TEI P3 Text Encoding Initiative. Revised reprint: Oxford

Wittenburg, P., Broeder, B. and Sloman, B. (2000) “International Standards for Language Engineering, Metadata Initiative (IMDI) White Paper”

http://www.mpi.nl/ISLE/documents/papers/white_paper_11.pdf (20 Apr. 07)