ISO-standard Metadata Descriptors and Registries

Lee Gillam

Department of Computing, School of Electronics and Physical Sciences, University of Surrey, Guildford, GU2 7XH, United Kingdom

l.gillam@surrey.ac.uk

Abstract. This paper discusses international standards (ISO) for metadata descriptors and registries that can be used to mediate data interchange and to improve data reuse – perhaps turning *data tombs* into *data wombs*. An approach to using standards-conformant metadata for creating interoperable markup formats is described that may have wider applicability to management and reuse of social sciences data, with potential for improvements in data fusion. The approach was developed in EU supported activites, emerged in a Terminology markup framework (TMF, ISO 16642), can be used both to document and to specify interoperable formats, and is being applied in for standardising other metadata and formats.

Introduction

The ability of scientists to examine each other's work by following similar steps to derive similar or different theorems, or by repeating experiments and observing similarities or differences in the results, is a cornerstone of all forms of scientific exploration and discovery. This entails the ability to share - or be able to use near-identical - tools, techniques, and data. Improvements to sharing social sciences data using standardized metadata could be a basis for improving the tools with which social sciences research is carried out and for scaling-up social sciences research: better management of data, growing at a rate of a few exabytes (billion gigabytes) of unique information each year, is required longer term before large-scale resources can be used. With increasingly inexpensive digital devices such as voice recorders, and an almost ubiquitous coverage of CCTV, we can only expect increases in the scale of collections of digital resources. Commensurate increases in the ability to manage and analyse such collections are also required. Standardisation of metadata enables the convergence of mark-up formats and improves the reusability of data. One of the successes of the eXtensible Markup Language (XML) is the ease with which it can be used to produce interchange formats: the resulting profusion of interchange formats necessitates interchange between interchange formats.

This paper discusses ISO-conformant metadata descriptors for the mark-up and interchange of language resources such as terminologies. An example is given for deriving a standards-conformant format that can be used to mark-up and exchange domain-specific terms that uses the combination of a *metamodel* from ISO 16642 and a set of metadata descriptors that conform to ISO 12620 ("Data Elements" stored in metadata registries, as defined in ISO 11179-3). Collections of such terms may be useful elsewhere as a basis for *ontologies* (Gillam, Tariq and Ahmad 2005, Gillam 2004). An ISO conformant metadata registry that supports such operations, and may be more broadly applicable, is currently in use in the EU eContent project Linguistic Infrastructure for Interoperable Resources and Systems (LIRICS).

Grids, metadata and markup

Foster and Kesselman identify the development of standards as "an important step toward the realization [of the grid]" (1999: 29). These standards are important for provenance, quality and validity, "correct" combinations of metadata, alternate representations, personalisation and interchange (*ibid.* 105-6, 123-7). Uptake of ISO standards often appears to be ignored, intentionally or not, in Grid initiatives, perhaps due to inertia, unfamiliarity, or perception that updates to ISO standards cannot keep pace with technology.

Work on metadata for language resources has been supported by the EU under the IST programme (for example, Wittenburg, Broeder and Sloman, 2000). A parallel in the US is the American Open Language Archives Community (OLAC) initiative (Bird and Simons 2000). More international activities on metadata for language resources include the Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard 1999) and the XML Corpus Encoding Standard (XCES) (Ide, Bonhomme and Romary 2000). Interchange formats for various purposes increasingly use XML. XML-based interchange formats along with systems of metadata are intended to make "knowledge", in its different forms and structures, more "accessible" (the divide between knowledge and information is frequently unclear in such assertions). Metadata such as those in Dublin Core do not provide sufficient granularity for specific activities: insufficient granularity results in what some refer to as "tag abuse", and which others actively campaign against. The CCLRC Scientific Metadata Model and the Scientific Annotation Middleware activities are both concerned with managing experimental data, but do not deal with metadata at large; metadata has been identified as a "major gap" in in e-Science at large (e-Science Gap Analysis, 2003, p10).

ISO standards-conformant formats

EU supported terminology interchange initiatives resulted in the creation of a terminological markup framework (TMF, ISO 16642) (Gillam et al 2002). ISO 16642 specifies a model of models (metamodel) that emerged for terminology markup. The metamodel abstracts away from specific database 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) (ISO 16642, p12).

The metamodel can be composed into a format by attaching metadata elements to its nodes. The metamodel is used in combination with metadata descriptors described in ISO 12620 to produce interoperable and reusable terminology resources.

For example, to specify a (minimal) XML format for terminology interchange we: (i) instantiate the metamodel from (ISO 16642) to produce the structure {e.g. TE [LS [TS]]]} (ii) select metadata elements from (ISO 12620) {e.g. *Term; language identifier; definition*} (iii) anchor metadata elements to the structure as (ISO 16642) {e.g. TE {*definition*}}; (iv) provide *style* and *vocabulary* to produce XML (ISO 16642) {e.g. **style**: "language identifier as XML Attribute; rest as XML Elements; **vocabulary**: definition = "def"; language identifier = "lang"; term = "term"}. A fragment of such an XML format would look something like:

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

```
</TE>
```

A similar approach can be used to describe existing formats by mapping to the metadata and the structural skeleton, and enables the degree of interoperability between formats intended for the same purpose to be determined. As such, interchange formats can be composed that are less "Anglo-Saxon", or that discourage human readability and interpretation – XML is, after all, intended for machine-processing.

Discussion

Arguably, the stability offered by ISO standards is essential for widespread uptake of Grid technologies. ISO standards such as these may eventually underpin Grid initiatives, specifically with reference to Semantic Grids (de Roure, Jennings and Shadbolt 2003) and emerging Knowledge Grids (Cannataro and Talia 2004). The cost of purchasing ISO standards and the perceived impenetrability of their contents may be perceived as barriers to their widespread use. It is worth considering, however, that existing international standards of varying applicability include those for: standard quantities (ISO 31-0); language codes (ISO 639-1, ISO 639-2); country codes (ISO 3166); dates (ISO 8601); character sets (ISO 10646) and data types (ISO 11404). Adopting and defining standards-conformant metadata has the potential, longer-term, for improving management of data and facilitating the scaling up of research – perhaps even helping to automate complex data preparation and analysis routines and facilitate new forms of collaborative research. Research results identified using finer-grained metadata become available for use beyond that originally intended, ethical considerations allowing, although additional efforts are required to produce these metadata. When large quantities of data are "deposited to merely rest in peace, since in all likelihood it will never be accessed again", the phrase **Data Tombs** is used (Fayyad and Uthurusamy 2002). Use of standardised metadata to promote reusability could result in the emergence of **Data Wombs** where data collections that will evolve over time are sustained and nurtured.

The metamodel approach is being applied to the standardisation of other language resources, including language identifiers (Dalby and Gillam 2004), and is needed to counter the profusion of markup formats for identical, or highly related purposes. Selection of a metamodel for a given purpose, combined with both purpose-specific and more generic metadata, can be used to produce an interchange format, demonstrated here for terminology formats. Efforts are needed to define and disseminate the use of such sets of metadata, contributing to and benefiting from activities both in the UK through the British Standards Institution (BSI) for identification of languages, and through involvement with ISO for language resource metadata. Production of a metadata registry for syntactic and morphological annotation and representations of semantic content are the subject of a current EU e-Content project called LIRICS. A similar effort may be worthwhile considering for a metadata registry for social sciences. Such an effort could contribute to overcoming a need for "many non-Grid metadata organizations whose repositories need to be wrapped" (ibid) and result in training courses that link XML training to data and related resources – metadata standards (Cole et al., e-Social Science scoping study). Work is planned that explores the use of standardised metadata in Data Grids, using the Storage Resource Broker (SRB) and its Metadata Catalog (MCAT).

Acknowledgment

This work was supported in part by the EU (SALT: IST-1999-10951, GIDA: IST-2000-31123, LIRICS: eContent-22236) and ESRC (FINGRID: RES-149-25-0028). The author is grateful to the reviewers of the original submission for their helpful 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 (17 Dec. 04)

Cannataro, M. and Talia, D. (2004). " Semantics and Knowledge Grids: Building the Next-Generation Grid". IEEE Int.Sys 19(1), pp56-63

Dalby, D. and Gillam, L. (2004) “Weaving the Linguasphere: LS 639, ISO 639 and ISO 12620”. Proc. of INTERA workshop, LREC 2004.

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 Fransisco, California.

Gillam, L., Tariq, M. & Ahmad, K. (2005). “Terminology and the Construction of Ontology?”, *Terminology* 11(1). John Benjamins, Amsterdam. (In Press).

Gillam, L. (2004). “*Systems of concepts and their extraction from text*”. Unpublished PhD thesis, University of Surrey.

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

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/IEC 11179-3:1994 Information technology -- Specification and standardi-zation of data elements. Part 3: Basic Attributes of Data Elements. 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.

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.

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 (17 Dec. 04)