Metadata, semantics, and ontology: providing meaning to information resources

Miguel-Ángel Sicilia

Computer Science Department,
University of Alcalá, Ctra. Barcelona, km. 33.6 – 28871
Alcalá de Henares, Madrid, Spain
E-mail: msicilia@uah.es

Abstract: *Metadata research* has emerged as a new discipline in the last years, and is focused on the provision of semantic descriptions of a diverse kind to digital resources, web resources being the most frequent target. Such associated descriptions are supposed to serve as a foundation for advanced, improved services in several application areas, including search and location, personalisation, and automated delivery of information. In consequence, metadata research focuses both on the development of metadata description languages – of a general purpose or specialised kind – and also on the practicalities of metadata creation, dissemination, assessment, maintenance, and use for diverse scenarios and usage contexts. *Ontology* has emerged recently as a knowledge representation infrastructure for the provision of shared semantics to metadata, which essentially forms the basis of the vision of the *Semantic Web*. The combination of metadata description techniques and ontology engineering defines a new landscape for information engineering with specific challenges and promising applications, which requires a truly multi-disciplinary approach. This paper is intended to provide some basic insights for the endeavour of engineering systems based on metadata, semantics, and ontologies, and to foster the interaction of researchers with different backgrounds coming from diverse disciplines.

Keywords: metadata; semantics; ontology.

Reference to this paper should be made as follows: Sicilia, M-Á. (2006) 'Metadata, semantics, and ontology: providing meaning to information resources', *Int. J. Metadata*, *Semantics and Ontologies*, Vol. 1, No. 1, pp.83–86.

Biographical notes: Miguel-Ángel Sicilia obtained his degree in Computer Science from the Pontifical University of Salamanca, Madrid, Spain in 1996 and his PhD from Carlos III University in 1999. From 1997 to 1999 he worked as an Assistant Professor and later as a part-time Lecturer at the Computer Science Department of the same university. He also worked as a Software Architect in e-commerce consulting firms. From 2002 to 2003, he worked as a full-time Lecturer at Carlos III University, after which he joined the University of Alcalá. His research interests are primarily in the areas of adaptive hypermedia, learning technology, and human-computer interaction, with a special focus on the role of uncertainty and imprecision handling techniques in those fields.

1 Introduction

Metadata is today the subject of much research and debate, especially as a result of the increasing popularity of the web. This has resulted in several relevant standardisation efforts, including the *Dublin Core* Initiative (DCI)[1] and other general-purpose schemas, and also in a plethora of proposed specifications for a wide range of domains or application areas. Examples are the recent HrXML specifications, covering human-resource management data interchange, and also the multiple e-learning standards and specifications.

The most common definition for metadata says, 'Metadata is data *about* data'. But this generic definition does not capture the richness of possibilities of description of digital resources. Greenberg (2003) defines metadata as "*structured* data about an object that supports *functions* associated with the designated object", introducing two important aspects. Structure in metadata entails that information is organised systematically, and this is an aspect that is far from being controversial, especially due to the fact that metadata for many domains is nowadays subject to standardisation. The term metadata schema is often used to refer to one such specific organisation. Nonetheless, the fact that metadata is created to support some specific function is sometimes overlooked or vaguely acknowledged. Even though some functions are tacit in metadata, e.g., a 'subject' metadata element is obviously intended for the function of discovery, or 'cost' is intended for a purchase activity, metadata creators are often not concerned with the concrete details of the requirements of the functions that will make use of the metadata records they generate. This is a problem of semantics in general, which is especially important when developing software that automatically process and reacts on metadata.

Metadata can be expressed in a diverse range of human and artificial languages and forms. In fact, even simple common HTML pages provide some embedded forms of metadata, e.g., data can be put in the title element, and we can even consider that annotations in the margin of a printed book are in fact a form of metadata. This raises the need for clarification about which kind of entities are subject to be referenced by metadata, and also what forms of metadata are possible or useful for particular needs. Then we need a notion of *resource*, and also a notion of *metadata language*. Resources will be the described objects, considering that metadata itself can be considered a resource subject to description. Metadata languages are the shared description systems used for the codification of metadata. Such systems vary in their expresiveness, but also in their room for ambiguity and imprecision. Natural languages provide a large degree of expresiveness to specify metadata, but are subject to ambiguity, and are, thus, not a good framework for processing by software systems.

Metadata has been proposed as a mechanism for expressing the 'semantics' of information, as a means to facilitate information seeking, retrieval, understanding, and use. But meaning can be considered as a 'locally constructed' artefact, as described by Brasethvik (1998), so that some form of agreement is required to maintain a common space of understanding. In consequence, metadata languages require shared representations of knowledge as the basic vocabulary from which metadata statements can be asserted. Ontology as considered in modern knowledge engineering (Gruber, 1993) is precisely intended to convey that kind of shared understanding. In consequence, ontology along with (carefully designed) metadata languages can be considered as the foundation for a new landscape of information management. Nonetheless, such an endeavour for 'semantic systems' requires the coincidence of the effort of many scientific, engineering, and management disciplines to become a reality. Such disciplines not only include artificial intelligence and knowledge representation, but also many others like library science, philosophy, education, database management, etc.

In the rest of this paper, some insights are provided about metadata research as an emerging discipline, shaped by the shared support for understanding provided by ontology, with the aim of providing some initial directions for an integrative view of metadata, semantics, and ontology as the foundations for better information systems.

2 Metadata

The main characteristic of metadata is its *referential* nature, i.e., metadata predicates about some other thing. Such 'other thing' can be considered as 'anything' from the broadest perspective, but such a view could hardly be useful for bringing semantics to current information systems as the web. Then, we will restrict our discussion to digital resources of a diverse kind. In the scope of the current web, resources can be unambiguously identified by the concept of URI, which is now able to address even fragments of mark-up languages by means of the *XPath* and *XPointer* syntaxes. This concept of resource is not restrictive to considering people or other entities as resources for particular purposes. For example, in educational settings, a tutor as an educational resource can be represented by some digital surrogate like an address or contact method, and the same is true with reference to locations for physical books inside libraries. Even concepts or conceptual works (e.g., 'Hamlet' as a work of art) could be represented by digital surrogates, e.g., as ontology instances – provided that some kind of agreement is set about such representations.

Then, a metadata fragment or piece is a statement about a (digital) resource. The fact that metadata is also represented in digital form and uniquely identified leads to the inherent capability of providing metadata about metadata. This recursive character of metadata is the origin of the notion of metadata *levels*, so that given a collection of information entities we can classify them in diverse meta-levels, some of them being primitive, i.e., those not referring to others. This recursive capability is consistent with flexible metadata representation languages like RDF, although some balance between richness and usability should be achieved in practical settings.

Metadata can be also used in at least two additional idiomatic variants, which complicate the provision of overall definitions. Metadata can be used to describe *relationships* between resources, and also to describe prospective *usages* of resources. In the relational idiom, metadata is a statement that says something about more than one resource. Links considered as independent to contents in hypermedia are the most common vehicle for this kind of metadata. But such links should carry some semantic description to be interpretable by software, e.g., link types (Trigg and Weiser, 1986) indicating the purpose of the connection, e.g., 'criticises', 'provides an analogy', etc. The use-oriented metadata idiom entails statement about the properties of the resource as used in a particular context. This is actually the case of educational metadata, in which educational contents are considered as having some properties when used in a concrete educational situation.

From these basic definitions, some important research questions can be formulated, including the following:

- which are the appropriate languages for describing metadata?
- what are the (right) vocabularies for metadata? how could they be standardised?
- how can metadata quality be assessed?
- how metadata inconsistencies can be handled?
- which tools are needed to efficiently produce high-quality metadata?
- which tools are needed to manage high volumes of metadata?
- which metadata management practices are required?

The quest for appropriate metadata languages has been a fundamental subject in Semantic Web research (Berners-Lee et al., 2001), resulting in web-ready languages like OWL that provide powerful capabilities in terms of their underlying description logics (Baader et al., 2003). But such logical foundation is not enough by itself, since the ways of using such languages should also be investigated. In other words, with OWL, we can make arbitrary assertions, but some agreement about their usage is required to obtain consistent and shared semantics. Ontologies engineered with such languages are considered as an answer for the second question about vocabularies, but once again, the practices of metadata creation are to a large extent a matter of standardisation and investigation about the practicalities and management aspects of information.

Moreover, the provision of an appropriate language by no means guarantees that the resulting metadata statements would be adequate as descriptions. This raises the need for quality assessment and quality certification mechanisms, especially for metadata created in open environments like the web. Such metadata quality problem is actually multi-faceted, since it requires at least new notions of completeness, trust, and consistency. *Completeness* of metadata is a concept oriented to state which metadata statements (or elements) are required for a particular kind of resources to be usable for some kind of process. For example, metadata about the cost of a resource is required for automated acquisition and trading. Honesty is a well-known problem in the web, and it is concerned with delimiting which information is actually provided for the sake of description, and which others have spurious interests, like the misuses of the meta tags in HTML to attract visitors. *Consistency* of metadata profiles is an even more challenging concept, since it is concerned with how to reconcile metadata statements that are somewhat contradictory. Concepts of authority, certification or perhaps collaborative filtering should be elaborated to advance in research about consistency of metadata.

Since metadata is intended to describe digital resources, it is reasonable to expect a huge growth in the amount of metadata in the following years, since the growth of the web also seems to continue. This 'volume' effect raises a sort of different questions for metadata research. Such questions include which tools should be required for creating and managing metadata in an efficient and effective way, and also if current institutional practices in libraries, universities, and other organisations are appropriate for the management of that growing amount of metadata.

In addition, all the previous questions can be subject to reformulation in the context of specific disciplines. For example, the vocabularies for describing pedagogical uses of learning objects are subject to intense debate (Allert, 2004).

3 Ontology and metadata research

Ontology is usually defined as (shared) specifications of conceptualisations (Gruber, 1993). Modern formal ontology has a dual nature. On the one hand, they are complex knowledge representation artefacts intended for the development of intelligent applications. But on the other hand, they are social constructions intended for communication and crystallisation of domain-specific knowledge. As such, they are subject to the evolution of disciplines or domains, and also to divergent or non-orthodox views of a field. Nonetheless, the capabilities of formal ontology to convey relationships and axioms make them an ideal vehicle for describing the vocabulary for metadata statements, providing a rich formal semantic structure for their interpretation. Several questions arise from the use of ontology as vocabularies for metadata, including the following:

- how should ontology be used in metadata statements?
- what level of ontological description is useful for the practical purposes of metadata?
- how could ontology be used to improve the design of the interaction and to express information needs?

The first question in the above list is often referred to as ontology *annotation*, and in practical terms, it requires a clarification of how metadata standards and schemas should make use of ontology, and what are the consequences of using ontological structures in services and systems. The second question is a re-formulation of the problem of usability of ontological representations that has been studied elsewhere (Russ et al., 1999). Ontology may also play a role in the design of human interaction, as advanced in some ontology-based seeking interfaces (García and Sicilia, 2003). These kinds of user interface approaches are also a new challenge for metadata, since their functioning is closely tied to metadata statements.

Attention should be given also to the concept of 'semantics' as related to ontology. As recently pointed out by Seth et al. (2004) formal ontology as usually considered in current Semantic Web applications can be complemented by 'soft' representations (e.g., fuzzy or possibility logics), and of course with the semantics that are implicit to the texts or media that actually conform the web.

In addition, much effort still remains in the crafting of ontology for many domains, in spite of the fact that many relevant ontological efforts are yet completed.

4 Conclusions

Metadata research can be considered as a multi-disciplinary field oriented to the development of improved technologies mediated by referential descriptions of resources in languages that are adequate for machine consumption. According to Mitcham (1994), the usefulness of any technology in any field is dependent on its capacity to address real problems and address practical needs in that field. This consideration is especially important in metadata research, since its evolution and success requires a huge amount of work in terms of providing rich metadata to existing resources. And such effort needs of course a value

justification that motivates individuals and organisations to engage in the crafting of semantic information systems. Some reflections about the nature of (digital) metadata have been provided in this paper, with the intention of fostering discussion and further clarification in the discipline of metadata research.

Nonetheless, the vision provided in this paper does not exhaust the range of research issues concerned with metadata, semantics, and ontology, and surely new challenges will appear with the evolution of technologies and practices of metadata creation.

References

Allert, H. (2004) 'Coherent social systems for learning: an approach for contextualized and community-centred metadata', *Journal of Interactive Media in Education*, Vol. 2, http://www-jime.open.ac.uk/2004/2.

Baader, F., Calvanese, D., McGuinness, D., Nardi, D. and Patel-Schneider, P. (Eds.) (2003) *The Description Logic Handbook. Theory, Implementation and Applications*, Cambridge.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) 'The semantic web', *Scientific American*, Vol. 284, No. 5, pp.34–43.

Brasethvik, T. (1998) 'A semantic modeling approach to metadata', *Internet Research: Electronic Networking Applications and Policy*, Vol. 8, No. 5, pp.377–386.

García, E. and Sicilia, M.A. (2003) 'User interface tactics in ontology-based information seeking', *Psychnology E-journal*, Vol. 1, No. 3, pp.243–256.

Greenberg, J. (2003) 'Metadata and the world wide web', in Dekker, M. (Ed.): *Encyclopaedia of Library and Information Science*, pp.1876–1888.

Gruber, T.R. (1993) 'A translation approach to portable ontologies', *Knowledge Acquisition*, Vol. 5, No. 2, pp.199–220.

Mitcham, C. (1994) *Thinking through Technology: The Path between Engineering and Philosophy*, The University of Chicago Press, Chicago.

Russ, T., Valente, A., MacGregor, R. and Swartout, W. (1999) 'Practical experiences in trading off ontology usability and reusability', *Proceedings of the Twelfth Banff Knowledge Acquisition for Knowledge-based Systems Workshop*, October 16–21, Banff, Alberta, Canada.

Seth, A. *et al.* (2005) 'Semantics for the semantic web: the implicit, the formal and the powerful', *Intl. Journal on Semantic Web and Information Systems*, Vol. 1, No. 1, pp.1–18.

Trigg, R.H. and Weiser, M. (1986) 'TEXTNET: a network-based approach to text handling', *ACM Transactions on Office Information Systems*, Vol. 4, No. 1, pp.1–23.

Note

[1]http://dublincore.org/.