Between symbol and language-in-use

Emma Tonkin

UKOLN,
University of Bath, BA27AY, UK
e.tonkin@ukoln.ac.uk
http://www.ukoln.ac.uk/

Abstract. Indexing is often designed with the intent of dimensional reduction, that is, of generating standardised and uniform descriptive metadata. This could be characterised as a process of decontextualisation. Formal knowledge representation systems typically have the aim of encapsulating granular pieces of information in a reusable manner. The result is a set of information elements with minimal links to external information sources. Plain-text tags, by comparison, have the aim of describing an object, within or outside a reductively described context. The result is a set of views that are contextualised to author, time, location, task or community. This paper discusses the relationship between symbol, contextual relation and language-in-use.

Key words: Language-in-use, metadata analysis, informal metadata

1 Introduction

Formal knowledge representation systems are often reductionist in philosophy; that is to say, a goal of techniques such as metadata collection is often the establishment of a dimensionally reduced set of data records from which to operate. In this sense, we might cast the process as one inherently concerned with decontextualisation, occurring in the Aristotelian tradition[5]. All extraneous variables are normalised to default values. Whether a domain specific encoding is used or a generic sub-language, the derived result is enforced homogeneity. The aim, though it is contentious to what extent the aim is practically realised, is generally to relate objects (physical, electronic or conceptual) to a formally defined model.

This paper is intended to contribute to an existing discussion regarding the process of developing a formal system and the gradient between natural language, metadata and formal system. This is of relevance to a number of topics in information retrieval. Formalisation in knowledge representation and retrieval is a mature topic. Wilks (2006) discusses the relationship between natural language and formalism in terms of the Semantic Web, describing 'two differing lines of Semantic Web research: one, closely allied to notions of documents and natural language (NL) and one not.' Here, we ask a similar question in terms of semi-formal and informal metadata in use.

1.1 Representation in natural language

One might begin by asking the simplest of questions: What is a language? As a functional definition, a language allows one to speak and be understood by others who know that language [3]. It is possible to describe a language as a system composed of symbols with referential meaning; however, such a description does little to clarify that which is intended. Deacon [2] articulates two definitions of *symbol*, one drawn from the humanities and another drawn from computation.
Humanities: A symbol is one of a conventional set of tokens that marks a node in a complex web of interdependent referential relationships and whose specific reference is not obviously discernible from its token features. Its reference is often obscure, abstract, multifaceted and cryptic, and tends to require considerable experience or training to interpret.
Computation: A symbol is one of a conventional set of tokens manipulated with respect to certain of its physical characteristics by a set of substitution, elimination and combination rules, and which is arbitrarily correlated with some referent.

To precise further, then, would place this paper on one side or another of an interdisciplinary rift - a dichotomy both apparent and keenly felt. What is the validity of this voluntary separation between disciplines?

Deacon characterises these uses as 'complementary, referring to two different aspects of the same phenomenon', with the computational definition predominantly describing the production and manipulation of the symbol tokens and the former definition relating rather to the interpretation and symbolic effects of symbolic reference. The distinction is drawn between icon and system – an icon could be said to form a reference by drawing on similarity to a referent, a resemblance, whilst a symbolic token has no such constraints, though it may act as an icon in certain contexts. However, the relevance of iconicity in acquisition of a symbol is questionable (see [8], p.36-37). Links between symbol and world are drawn by indexical reference - a mapping between symbol and referent, based on correlation.

1.2 Natural language to formal representation

There exist formal treatments of natural languages, characterising NL as a set of algebraic rules and a lexicon of meaningful linguistic elements[6]. There exist also usage-based theories of linguistics that treat structure as resultant (emergent) from language use [8].

One could place various approaches toward information representation on a spectrum of increasing regularity or completeness of intended definition of symbol and/or relation. Alternatively, one could order systems according to the set of assumptions which underlie each approach; for example, some simple statistical

systems rely simply on term frequency/keyword density, with others depending instead on the distributional hypothesis. Metadata schemas are often designed explicitly for use within a single environment, with a more or less completely defined set of use cases. Context is explicitly handled by metadata application profiles, which provide a means of labelling records as resulting from a given application type of a metadata schema.

In practice, a representation designed for information management and retrieval purposes is typically influenced by concerns other than cognitive or neural realism. Computability, for example, is a primary concern. An optimal representation may therefore be far from realistic. From the HCI viewpoint, a probable approach for creating such a representation is likely to make use of information elicited from study of appropriate stakeholders. An alternative approach is the encoding and use of an existing formal representation. Either way, a formal classification is far from decontextualised; to quote Stephen Jay Gould on the purpose of classification: "[Classifications represent] theories about the basis of natural order, not dull categories compiled only to avoid chaos."

The development of a formal representation involves decontextualisation in the sense that extraneous contextual information is given only implicitly. The context of a formal ontology is formally given only in the sense that a namespace is provided; the syntax rules and lexicon are provided with the implicit datum that they apply only within the operative context of this representation. However, this contextual information is not explicitly encoded – which statement is not intended to suggest that it is possible or desirable to do otherwise. The purpose of the process of formalisation is generally reductionist. The eventual aim is the extraction of information in a form usable for the diverse purposes of the system, which implies the need to collect information in a form appropriate for that purpose.

Deacon[2] notes Frege's recognition that 'words on their own generally do not refer to particular concrete things in this world except when in certain combinations or contexts that determine this link'. Brief utterances require explicit context to be appropriately interpreted. What explicitly given context has a key-value pair in, for example, Dublin Core metadata?

Applied language – language-in-use – acquires a syntax and semantics characteristic of its domain of use and of the actors between which the term is used. The design of formal systems for each knowledge subdomain or scenario represents a formal (analytical) approach to the same representation task that language systems in general approach in a more general manner. This leads us to a question that might be described as a recurring theme in digital library research: why, in a given scenario for metadata use, would we expect a formal system to be more "appropriate" than a system developed by participants in the process for use in the area – and by what metrics might we measure appropriateness?

2 Metadata as language-in-use

An approach toward formal representation of information is no more accurate than the users who apply it. Wilks[9] points out that this is a standard philosophical problem; as annotations are used to bind text to meaning representations, the markers themselves are said by some critics to take up the characteristics of natural language and therefore reach no meaning outsid language. As a solution, linking the virtual world to real-world quantities and artifacts is suggested. Though accepted as a plausible approach, Wilks adds that 'Nothing will satisfy a critic... except a web based on a firm (ie. formal and extra-symbolic) semantics and effectively unrelated to language at all [...] The SW may be the best way of showing that a non-formal semantics can work effectively, just as language itself does and in the same way.'

This is the most revealing of quotes. If designers of informal and semi-formal semantic systems are building languages, then that language may be expected to be as susceptible to contextualisation in speech acts as any other. Is it possible that appropriate analysis of real-world use of existing indexing systems (in the sense of annotation systems rather than textual analysis approaches - although application of such techniques may well qualify as 'appropriate analysis') would show that contextualised use of metadata is already with is, though the encoding is not an explicit one?

2.1 Tagging systems

Plain-text tags have the aim of describing an object and providing a pointer to that object - the generation and use of free-text metadata for description and discovery of resources. The result is a set of views that are contextualised to author, time, location, task or community. The tag is the vaguest of indexing systems. A tag corpus is constructed of a set of speech acts, and each term is generally devoid of context in the sense of grammar or syntactic relatives. The relative of the distributional hypothesis in tagging could better be labelled the " co-occurrence hypothesis" - similar words are preferentially used to point to similar items.

Tags are simply snatches of natural language, though some efforts have been made to encourage consistent use of conventions such as spelling, pluralisation and so forth. There is an argument to be made that tags are simply keywords, and indeed the difference is more likely to be found in the domain of use and characteristics of the user community than in the technology itself. Either way, tag corpuses provide a fascinating opportunity to examine a largely user-driven adaption of natural language for indexing purposes. Any reductionist influence present in this subset of language exists either due to technical limitations or the decision of the individual providing the tag. This provides for the fascinating possibility that a limited subsystem of language can arise from applied use of

natural language in a given context – a folksonomy. The characteristics of such a sublanguage are in general studied, rather than as a corpus of interest to linguistics, as a keyword corpus in need of filtering.

2.2 Evidence from semi-formal metadata

It is undoubtedly easy to point to patterns of failure in the application of semi-formal metadata systems, such as for example Dublin Core application profiles in a given information retrieval context. However, to pinpoint the causes of such failures is relatively difficult. It is probable that they frequently result from problems such as ambiguity in key names as interpreted by the user community – that is, misunderstanding of the intended use of a given field – and changes in the scope of use of a given schema following its introduction.

One might describe this, somewhat flippantly, as analogous to the Whorfian hypothesis in action. Where the hypothesis suggests that one cannot think something for which one does not possess a linguistic representation, this relates instead to the assumption that one cannot represent something for which one does not have an appropriate element in one's schema. This, of course, is false; in practice user populations typically manage very well in the face of unexpected requirements, sacrificing interoperability by applying a sensible-sounding self-sponsored adaption to the system. Such adaptions may be characterised according to many factors. Drivers such as intended audience and convenience are significant. What prompts such inventions, and under what circumstances does the motivation for incorporating an original concept overcome deterrent factors?

Tennis[7] notes that tagging "seems intensely personal, whereas subject cataloguing is an act of delegation mediated by institutions," drawing a clear distinction between indexing as a prescriptive and as a descriptive process. Both processes take place in a definable context - in the first, the context is personal ; in the second, it is institutional. The intended audience of descriptive speech has a significant impact on the ease by which it may be interpreted after the fact (see for example the experiment described by Lave[4]). With this and the earlier discussion of classification as theory in mind, it seems appropriate to ask whether the construction of many current information retrieval systems does not already amount to a set of suppositions regarding the context of use.

To come to an understanding of the domains in which an information retrieval system succeeds or fails is a special case of a general problem; that of the appropriateness of a symbolic system for a given case. The handling of context in natural language itself is far from simple, though it may be modelled in a number of ways, such as by application of variants on the distributional hypothesis. Language in general carries various indicators of context on syntactic and semantic levels. It is reasonable to expect that in practical application, formal symbolic

systems will acquire (and very probably already exhibit) a very similar character.

3 Conclusion

The possibility of formally encoding the notion of context inspires a counterpoint question – is it possible to formally exclude the notion? The capacity for creating a formal representation does not necessarily imply that such a representation is wanted or needed; there is an argument to be made that arbitrarily created representations are theories, ways of classifying the world around us. In many cases, it is likely that the need for the structures themselves is not as yet ascertained.

The design of information retrieval systems is complicated by a number of factors, one of which is the difficulty of establishing situationally appropriate metrics for evaluation. Ultimately, the question to be asked may be *why?* Natural languages can perhaps be characterised as compromising between a variety of competing aims, and artifically created or defined languages may be characterised similarly.

A current aim of our research is to examine existing corpora of informal and semi-formal metadata and, from this information, to characterise present patterns of use of these approaches. We find it probable that for our purposes, the simplest approach to contextualised metadata is to work as far as possible with the markers already present in indexing data. To examine the process of creation and use of an existing corpus of data may tell us more about what is already encoded or may be retrieved from the dataset – at the least, this approach may prove beneficial from a vocabulary management perspective.

References

1. Bearman, D., Trant, J. (2005), Social Terminology Enhancement through Vernacular Engagement Exploring Collaborative Annotation to Encourage Interaction with Museum Collections D-Lib. **11** 9, doi:10.1045/september2005-bearman
2. Deacon, T. W. (2003) UG and Semiotic Constraints. In: Language Evolution. Ed: Morten H. Christiansen and Simon Kirby. Oxford.
3. Fromkin, V., Rodman, R. (1993) An Introduction to Language, Fifth Edition. Harcourt Brace Jovanovich.
4. Lave, J. (1991), *Situating learning in communities of practice.* In: L.B. Resnick, J. Levine and S. D. Teasley (Eds.) Perspectives on socially shared cognition. Washington, DC: American Psychological Association.
5. Peterson, E. (2006), Beneath the Metadata: Some philosophical problems with folksonomy. D-Lib. **12** 11, doi:10.1045/november2006-peterson
6. Pinker, S. (1999), *Words and Rules: The Incredients of Language.* New York: HarperCollins.

7. Tennis, J.T. (2006) Social Tagging and the Next Steps for Indexing. In Furner, Jonathan and Tennis, Joseph T, Eds. Proceedings 17th SIG/CR Classification Research Workshop, Austin, Texas.
8. Tomasello, M. (2003), Constructing a language: A usage-based theory of language acquisition. Harvard University Press.
9. Wilks, Y. (2006) The Semantic Web as the apotheosis of annotation, but what are its semantics? International Journal of Web Semantics