Governance through Language: The Thesaurus as a Management Tool

++++++++++++++++++++++++++++++++++++++++++++++++

SHORT ABSTRACT

Taxonomies, thesauri and dictionaries have a long history as a means to capture knowledge and the language by which it is expressed. In organizational contexts, there is a growing interest in creating such artefacts. By doing so, organizations strive to attain a variety of related goals: capturing and dissemination of knowledge, normalization of terminology, and quality assurance — to mention a few. Increasingly, thesauri play a role in governing the design and operation of IT-systems. This is an interesting development, as it shifts the control over what data mean from the IT Department towards the business.

The shift results from a number of trends. Technology providers respond to these in innovative ways. Standards-compliant thesaurus management software offers an increasing number of features. This includes automatic "seeding" of a thesaurus by mining the DBPedia or large text corpora, but also mechanisms to govern "pick lists" such as "gender: {male, female, undisclosed}", so that software applications used in business processes literally use the thesaurus as the single source of reference data. An exciting development is adding features that leverage semantic integration through Linked Data technologies.

We will present an analysis of these trends. We also reflect on the different strategies to deal with polysemy, idiosyncrasies, and multilingualism and to organize governance. Finally, we discuss a real-world case.

++++++++++++++++++++++++++++++++++++++++++++++++

DESCRIPTION

The problem addressed in this presentation is ultimately a problem of governance.

As IT systems become larger, more complex and more important, assuring the quality of the information inside them starts to become an issue. Managing definition of terms through data dictionaries containing tens or hundreds of thousands of data items simply doesn't work. Moreover, it is the IT-department that, often unwillingly, is responsible for ownership. The alternative is a thematically organized thesaurus with a manageable amount of concepts, with broadly used and standardized methods to capture semantic relations between terms, and to deal with multilingualism, synonymy and homonymy. A group of executives within the business can then be formed to act as owner and regain control over what the information inside the IT-systems means. This is a prerequisite for managing data quality.

A second trend is the push for large scale, cross-border integration of IT systems, or rather, of the information contained in them. Technological interoperability standards, such a SOAP and REST, have enabled this push. At the same time, however, interoperability at the semantic level is a matter of human language. The same sociolinguistic principles that govern human interaction are at work. This means, for instance, that forcing shared data dictionaries on many organizations in a network in a top-down manner will meet impedance. Yet, this is the classic paradigm for creating interoperability standards. Modern approaches have learned from the rise of the Web: bottom up, allowing for mismatches by defining mechanisms to deal with these. What is needed, however, is allowance for "pidginization" and "creolization"

— that is, a language that develops in interaction rather than one being designed by an authority.

The innovative nature of this approach is the application of long-known techniques and methods from the library sciences in the domain of IT-governance. This application poses new challenges which can be elegantly met using Linked Data techniques.

Technology providers respond with creative solutions. Standards-compliant thesaurus management software offers an increasing number of features beyond what is needed for the traditional application of indexing libraries. This includes relatively simple but highly effective mechanisms to govern "pick lists" such as "gender: {male, female, undisclosed}", so that software applications used in business processes literally use the thesaurus as the single source of reference data. An exciting development is adding features that leverage semantic integration through Linked Data technologies for instance for the purpose of automatic "seeding" a thesaurus by mining the DBPedia or large text corpora.

A recent example of the latter development is an open thesaurus for legal terms published by Wolters Kluwer Germany (http://vocabulary.wolterskluwer.de), which is a controlled vocabulary enriched with information from DBPedia and Eurovoc.

The so-called "justitiethesaurus" is an example of an important, authoritative thesaurus in the justice domain in the Netherlands. Originally developed for the purpose of information retrieval in the research library published by the WODC, the research bureau of the Department of Security and Justice, it is currently used in modified and enriched form by several other organizations for their libraries. Increasingly, the thesaurus is looked at with an eye to use it as a basis for term management in IT-systems.

In the world of library management, the governance is often well-organized. The example of the "justitiethesaurus" shows how different organizations can cooperate effectively and reuse parts of each other's vocabulary, with enough flexibility to avoid paralysis. In the world of structured data, this type of cooperation is still perceived as challenging.

The central theme of this presentation is governance over the vocabulary through which the operation of an organization is manifested. Based on several examples, lessons learned will be drawn concerning strategies to organize this type of governance. On the one hand, a thesaurus must be authoritative so that it offers some measure of stability and quality. On the other, this authority must be localized enough so as to allow for effective and efficient decision making.

++++++++++++++++++++++++++++++++++++++++++++++++

LONG ABSTRACT

Taxonomies, thesauri and dictionaries have a long history as a means to capture knowledge and the language by which it is expressed. In organizational contexts, there is a growing interest in creating such artefacts. By doing so, organizations strive to attain a variety of related goals: capturing and dissemination of knowledge, normalization of terminology, and quality assurance — to mention a few. Increasingly, thesauri play a role in governing the

design and operation of IT-systems. This is an interesting development, as it shifts the control over what data mean from the IT Department towards the business.

The shift results from a number of trends. As IT systems become larger, more complex and more important, assuring the quality of the information inside them starts to become an issue. Managing definition of terms through data dictionaries containing tens or hundreds of thousands of data items simply doesn't work. The alternative is a thematically organized thesaurus with a manageable amount of concepts, with broadly used and standardized methods to capture semantic relations between terms, and to deal with multilingualism, synonymy and homonymy.

A second trend is the push for large scale, cross-border integration of IT systems, or rather, of the information contained in them. Technological interoperability standards, such a SOAP and REST, have enabled this push. At the same time, however, interoperability at the semantic level is a matter of human language. The same sociolinguistic principles that govern human interaction are at work. This means, for instance, that forcing shared data dictionaries on many organizations in a network in a top-down manner will meet impedance. Yet, this is the classic paradigm for enabling data exchange. Modern approaches have learned from the rise of the Web: bottom up, allowing for mismatches by defining mechanisms to deal with these. What is needed, however, is allowance for "pidginization" and "creolization" — that is, a language that develops in interaction rather than one being designed by an authority.

Technology providers respond to these trends in innovative ways. Standards-compliant thesaurus management software offers an increasing number of features. This includes automatic "seeding" of a thesaurus by mining the DBPedia or large text corpora, but also relatively simple but highly effective mechanisms to govern "pick lists" such as "gender: {male, female, undisclosed}", so that software applications used in business processes literally use the thesaurus as the single source of reference data. An exciting development is adding features that leverage semantic integration through Linked Data technologies.

A recent example of the latter development is an open thesaurus for legal terms published by Wolters Kluwer Germany (http://vocabulary.wolterskluwer.de), which is a controlled vocabulary enriched with information from DBPedia and Eurovoc.

We will present an analysis of these and related trends. We also reflect on the different strategies to deal with polysemy, idiosyncrasies, and multilingualism. Finally, we discuss ways to organize the processes for creating and managing the vocabulary and their governance. Finally, we discuss a real-world case.