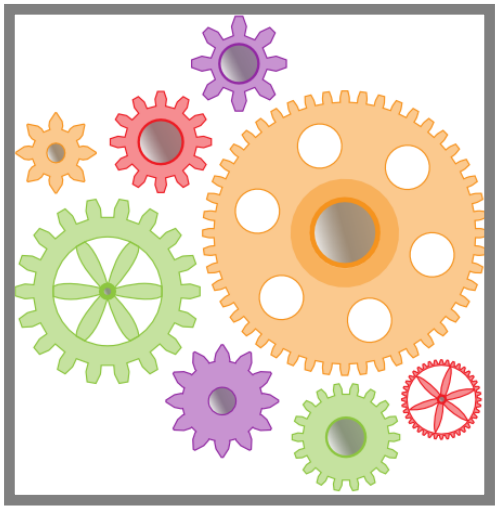
Making Text a First-Class Citizen

by Mike Bergman - Monday, January 28, 2013

http://www.mkbergman.com/1612/making-text-a-first-class-citizen/



Part 4 in the Enterprise-scale Semantic Systems Series

Text, text everywhere, but no information to link!

For at least a quarter of a century the amount of information within an enterprise embedded in text documents has been understood to be on the order of 80%; more recent estimates put that contribution at 90%. But, whatever the number, or no matter how you slice it, the percentage of information in documents has been overwhelming for enterprises.

The first documentation systems, Documentum being a notable pioneer, helped keep track of versions and characterized its document stores with some rather crude metadata. As document management systems evolved -- and enterprise search became a go-to application in its own right -- full-text indexing and search was added to characterize the document store. Search allowed better access and retrieval of those documents, but still kept documents as a separate information store from the true first citizens of information in enterprises -- structured databases.

That is now changing -- and fast. Particularly with semantic technologies, it is now possible to "tag" or characterize documents not only in terms of administrative and manually assigned tags, but with concepts and terminology appropriate to the enterprise domain.

Early systems tagged with taxonomies or thesauri of controlled vocabulary specific to the domain. Larger enterprises also often employ MDM (master data management) to help ensure that these vocabularies are germane across the enterprise. Yet, even still, such systems rarely interoperate with the enterprises' structured data assets.

Semantic technologies offer a huge leverage point to bridge these gaps. Being able to incorporate text as a first-class citizen into the enterprise's knowledge base is a major rationale for semantic technologies.

Explaining the Basis

Let's start with a couple of semantic givens. First, as I have explained many times on this blog, ontologies -- that is, knowledge graphs -- can capture the rich relationships between things for any given domain. Second, this structure can be more fully expressed via expanded synonyms, acronyms, alternative terms, alternative spellings and misspellings, all in multiple languages, to describe the concepts and things represented in this graph (a construct we have called "semsets".) That means that different people talking about the same thing with different terminology can communicate. This capability is an outcome from following SKOS-based best practices in ontology construction.

Then, we take these two semantic givens and stir in two further ingredients from NLP. We first prepare the unstructured document text with parsing and other standard text processing. These steps are also a precursor to search; they provide the means for natural language processing to obtain the "chunks" of information in documents as structured data. Then, using the ontologies with their expanded SKOS labels, we add the next ingredient of OBIE (ontology-based information extraction) to automatically "tag" candidate items in the source text.

Editors are presented these candidates to accept or not, plus to add others, in review interfaces as part of the workflow. The result is the final subject "tags" assignment. Because it is important to tag both subject concepts or named entities in the candidate text, Structured Dynamics calls this approach "scones". We have reusable structures and common terminology and syntax (irON) as canonical representations of these objects.

Add Conventional Metadata

Of course, not all descriptive information you would want to assign to a document is only what it is about. Much other structural information describing the document goes beyond what it is about.

Some of this information relates to what the document is: its size, its format, its encoding. Some of this information relates to provenance: who wrote it? who published it? when? when was it revised? And, some of this information relates to other descriptive relationships: where to download it? a picture of it; other formats of it. Of course, any additional information useful to describe the document can be also tagged on at this point.

This latter category is quite familiar to enterprise information architects. These metadata characterizations have been what is common for standard document management systems reaching back for three decades or more now.

So, naturally, this information has proven the test of time and also must have a pathway for getting assigned to documents. What is different is that all of this information can now be linked into a coherent knowledge graph of the domain.

Some Interface and Workflow Considerations

What we are seeking is a framework and workflow that naturally allows all exisitng and new documents to be presented through a pipeline that extends from authoring and review to metadata assignments. This workflow and the user interface screens associated with it are the more difficult aspects of the challenge. It is relatively straightforward to configure and set up a tagger (though, of course, better accuracy and suitability of the candidate tags can speed overall processing time). Making final assignments for subject tags from the candidates and then ensuring all other metadata are properly assigned can be either eased or impeded by the actual workflows and interfaces.

The trick to such semi-automatic processes is to get these steps right. There are the needs for manual overrides when the suggested, candidate tags are not right. Sometimes new terms and semset entries are found when reviewing the processed documents; these need to be entered and then placed into the overall domain graph structure as discovered. The process of working through steps on the tag processing screens should be natural and logical. Some activities benefit from very focused, bespoke functionality, rather than calling up a complicated or comprehensive app.

In enterprise settings these steps need to be recorded, subject to reviews and approvals, and with auditing capabilities should anything go awry. This means there needs to be a workflow engine underneath the entire system, recording steps and approvals and enabling things to be picked up at any intermediate, suspended point. These support requirements tend to be unique to each enterprise; thus, an underlying workflow system that can be readily modified and tailored -- perhaps through scripting or configuration interfaces -- is favored. Since Drupal is our standard content and user interface framework, we tend to favor workflow engines like State Machine over more narrow, out-of-the-box setups such as the Workflow module.

These screens and workflows are not integral to the actual semantic framework that governs tagging, but are essential complements to it. It is but another example of how the semantic technologies in an enterprise need to be embedded and integrated into a non-semantic environment (see the prior architecture piece in this series).

But, Also Some Caveats

Yet, what we have described above is the technology and process of assigning structured information to documents so that they can interoperate with other data in the enterprise. Once linked into the domain's knowledge graph and once characterized by the standard descriptive metadata, there is now the ability to search, slice, filter, navigate or discover text content just as if it were structured data. The semantic graph is the enabler of this integration.

Thus, the entire ability of this system to work derives from the graph structure itself. Creating, populating and maintaining these graph structures can be accomplished by users and subject matter experts from within the enterprise, but that requires new training and new skills. It is impossible to realize the benefits of semantic technologies without knowledgeable editors to maintain these structures. Because of its importance, a later part in this series deals directly with ontology management.

While ontology development and management are activities that do not require programming skills or any particular degrees, they do not happen by magic. Concepts need to be taught; tools need to be mastered; and responsibiilties need to be assigned and overseen to ensure the enterprise's needs are being met. It is exciting to see text become a first-class information citizen in the enterprise, but like any purposeful human activity, success ultimately depends on the people involved.

NOTE: This is part of an ongoing series on enterprise-scale semantic systems (ESSS), which has its own category on this blog. Simply click on that category link to see other articles in this series.
