The D-SCRIBE Process for Building a Scalable Ontology

Bob Schloss, Rosario Usceda-Sosa, Biplav Srivastava
IBM Research
{rschloss@us, rosariou@us, sbiplav@in}@.ibm.com

Abstract

In this paper, we describe the D-SCRIBE process used to build ontologies that are expected to have significant domain expansion after their initial introduction and whose coverage of concepts needs to be validated for a series of related applications. This process has been used to build SCRIBE, a very modular, ambitious ontology for the information about events triggered by both humans or nature, response activities by agencies that provide public services in cities by using resources and assets (land parcels, buildings, vehicles, equipment) and their communication (requests, work orders, sensor reports). SCRIBE reuses concepts from previously existing ontologies and data exchange standards, and D-SCRIBE retains traceability to these source influences.

Introduction

There are many approaches for building ontologies (Brusa et al 2006, Missikoff and Navigli 2005, Gruninger and Fox 2005). The most common is to extract domain knowledge and rules from experts and then encode them formally in an ontology representation. However, access to experts may not always be readily available, and it may become a costly and time-consuming process building consensus among multiple experts. A second approach is to bootstrap model acquisition by learning a probable model from secondary data sources likes plain text in documents (e.g., manuals, design documentation, web pages) or online resources over the web about a particular domain as captured in Word documents.

Most approaches look at usage scenarios at early stages. For example, in (Gruninger and Fox 2005), the authors start the process by interacting with the users and enquiring about the problem for which the ontology-enabled information system will be used. The process then uses the problems to specify the requirements in the form of questions that the developed ontology must be able to answer followed by defining the terminology of the ontology including its objects, attributes and relations. The next step is to specify the definitions and constraints on the terminology using logic. Finally, the completeness of the ontology is checked by testing it against the requirement questions.

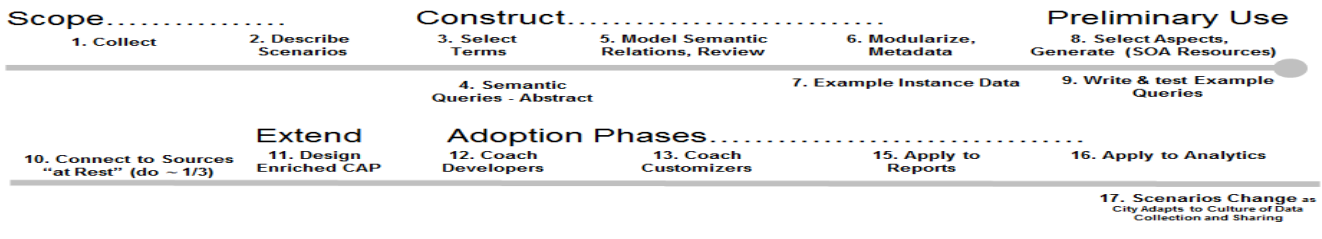
In a previous work (Uceda-Sosa et al 2011), we described the SCRIBE reference ontology centered around a set of semantic data models which are composable and customizable for different cities. In this paper, we describe the D-SCRIBE process used to build it efficiently with usage in focus.

The D-SCRIBE Process

D-SCRIBE process is a 16-step process from the stage of collecting the requirements for the ontology to using the models in applications and enhancing it based on usage. It relies on human domain expertise, where available, but also takes advantage of text extraction tools to discover concepts.

Step 1 is an inventory taking step about all the essential data concepts that real cities record and track, real software consumes and produces, and which is interchanged between city agencies, utilities, and businesses and individuals in the city. We looked at data from a handful of cities, a handful of software packages, but a dozen or more data exchange or data reporting standards. In our attempt to keep SCRIBE comprehensive but not "encyclopedic", we use a series of grounded scenarios, organized by different service obligations or expectations in cities, in step 2. To the extent that a candidate data item identified in Step 1 is not vital to any of our scenarios, it is dropped for now from inclusion in SCRIBE. Sometimes during step 2, a scenario collected from a client causes us to realize that there was a class of data exchange standards we had not considered, causing us to return back to step 1.

In Steps 3, and 4 we use information extraction tools to discover content from scenarios. Content extraction, and



the closely related area of information extraction, is a well-researched sub-area of information management with many research prototypes and commercial tools. A good survey for information extraction techniques is (Sarawagi 2007). The extracted concepts give us the potential concepts for modeling while the hierarchy serves as the basis for plausible relationships.

An important consideration for the created ontology for us is to promote consensus among potential users. In Step 4, we put the potential candidate concepts up for community review and then show links to how the concepts where used in the output ontology. At this point, all data concepts that we expect we will put in SCRIBE are added to a Data Dictionary which we hold in InfoSphere Business Glossary (IBM 2012), along with a business level definition, where possible a set of example values. Later this modularized dictionary (Steps 5 and 6) is delivered to our clients so that they can see exactly what any column in a produced reports means, in natural language. There are links in the individual glossary items back to the portion of the standard/source document, as well as, where appropriate, to matching entries in Wikipedia.

In Step 7, we actually add to our ontology instance data, because this confirms that what we have is expressive enough, and we find that others can learn how we intended our concepts to represent particular events or situations or dependencies by looking at concrete examples. Step 8 has to do with API generation, primarily for remote programs which will not access the ontology only through SPARQL or through the full Java API. In Step 9, the abstract queries and formulas that we developed in Step 4 are now expressed in an executable form (normally SPARQL) and we confirm that all the relations we intended to have in the model are present.

In Step 10, we work with a specific city, here called A-city, and their specific data sources, and do enough of the mapping so that there will be good models for the client to continue doing on their own. In Step 11, we consider the case that updates to city status are notifications to external parties are conveyed with the OASIS Common Alerting Protocol messages, extended with optional structured fields so that if software processes the message, additional precision and context can be conveyed. Steps 12-16 put this to use with real application developers, and augment the model for customization from A-city to any other city using a tool. Step 17 is an extension to handle changes.

We distinguish between application developers, who will write Java code and SPARQL queries, and customers, who will simply modify certain strings (for example, the name of the city) or augment certain enumerated code lists, or specify other customizations (such as provide a polygon showing city boundaries). Our future work will be to see how much the customization can be done by people who are not familiar with OWL modeling.

Conclusion

In this paper, we described the D-SCRIBE process to build a scalable ontology. It was used to create SCRIBE but it can also be used for building similar general-purpose ontologies that need to be used in different contexts by a diverse set of stakeholders. The unique aspects of the process are that it uses secondary sources (documents) to bootstrap identification of candidate concepts and relationships, allow community based review and tracking from candidates to actual concepts in the ontology, and allows selective customization of the ontology by different category of users.

References

Graciela Brusa, Ma. Laura Caliusco, Omar Chiotti, A Process for Building a Domain Ontology: an Experience in Developing a Government Budgetary Ontology, Australasian Ontology Wk. (AOW 2006), Hobart, Australia. At: http://129.96.12.107/confpapers/CRPITV72Brusa.pdf Accessed 8 Apr, 2012

Michele Missikoff, Roberto Navigli, 2005. Applying the Unified Process to Large-Scale Ontology Building, Available at: www.dsi.uniroma1.it/~navigli/.../IFAC_2005_Missikoff_Navigli.pdf Accessed 8 Apr, 2012

Rosario Uceda-Sosa, Biplav Srivastava, Robert J. Schloss, Building a Highly Consumable Semantic Model for Smarter Cities, IJCAI Workshop on AI for a Smarter Planet, Barcelona, 2011, Available at: http://dl.acm.org/citation.cfm?id=2018316

S. Sarawagi. Information extaction. In Foundations and Trends in Databases, Vol. 1, No. 3, Pg. 261 to 377, 2007.

IBM, Infosphere Business Glossary, 2012. At www.ibm.com/software/data/infosphere/business-glossary/ Accessed 19 Apr, 2012

Michael Gruninger and Mark Fox, Methodology for the Design and Evaluation of Ontologies, 2005.