Sources and Classification of Semantic Heterogeneities

by Mike Bergman - Tuesday, June 06, 2006

http://www.mkbergman.com/232/sources-and-classification-of-semantic-heterogeneities/

Semantic mediation -- that is, resolving semantic heterogeneities -- must address more than 40 discrete categories of potential mismatches from units of measure, terminology, language, and many others. These sources may derive from structure, domain, data or language.

Earlier postings in this recent series traced the progress in climbing the data federation pyramid to today's current emphasis on the semantic Web. Partially this series is aimed at disabusing the notion that data extensibility can arise simply by using the XML (eXtensible Markup Language) *data representation* protocol. As Stonebraker and Hellerstein correctly observe:

> *XML is sometimes marketed as the solution to the semantic heterogeneity problem . . . . Nothing could be further from the truth. Just because two people tag a data element as a salary does not mean that the two data elements are comparable. One could be salary after taxes in French francs including a lunch allowance, while the other could be salary before taxes in US dollars. Furthermore, if you call them "rubber gloves" and I call them "latex hand protectors", then XML will be useless in deciding that they are the same concept. Hence, the role of XML will be limited to providing the vocabulary in which common schemas can be constructed.[1]*

This series also covers the ontologies and the OWL language (written in XML) that now give us the means to understand and process these different domains and "world views" by machine. According to Natalya Noy, one of the principal researchers behind the Protégé development environment for ontologies and knowledge-based systems:

> *How are ontologies and the Semantic Web different from other forms of structured and semi-structured data, from database schemas to XML? Perhaps one of the main differences lies in their explicit formalization. If we make more of our assumptions explicit and able to be processed by machines, automatically or semi-automatically integrating the data will be easier. Here is another way to look at this: ontology languages have formal semantics, which makes building software agents that process them much easier, in the sense that their behavior is much more predictable (assuming they follow the specified explicit semantics--but at least there is something to follow).* [2]

Again, however, simply because OWL (or similar) languages now give us the means to represent an ontology, we still have the vexing challenge of how to resolve the differences between different "world views," even within the same domain. According to Alon Halevy:

> *When independent parties develop database schemas for the same domain, they will almost always be quite different from each other. These differences are referred to as semantic*

> *heterogeneity, which also appears in the presence of multiple XML documents, Web services, and ontologies--or more broadly, whenever there is more than one way to structure a body of data. The presence of semi-structured data exacerbates semantic heterogeneity, because semi-structured schemas are much more flexible to start with. For multiple data systems to cooperate with each other, they must understand each other's schemas. Without such understanding, the multitude of data sources amounts to a digital version of the Tower of Babel.* [3]

In the sections below, we describe the sources for how this heterogeneity arises and classify the many different types of heterogeneity. I then describe some broad approaches to overcoming these heterogeneities, though a subsequent post looks at that topic in more detail.

Causes and Sources of Semantic Heterogeneity

There are many potential circumstances where semantic heterogeneity may arise (partially from Halevy [3]):

- Enterprise information integration
- Querying and indexing the deep Web (which is a classic data federation problem in that there are literally tens to hundreds of thousands of separate Web databases) [4]
- Merchant catalog mapping
- Schema *v.* data heterogeneity
- Schema heterogeneity and semi-structured data.

Naturally, there will always be differences in how differing authors or sponsors create their own particular "world view," which, if transmitted in XML or expressed through an ontology language such as OWL may also result in differences based on expression or syntax. Indeed, the ease of conveying these schemas as semi-structured XML, RDF or OWL is in and of itself a source of potential expression heterogeneities. There are also other sources in simple schema use and versioning that can create mismatches [3]. Thus, possible drivers in semantic mismatches can occur from world view, perspective, syntax, structure and versioning and timing:

- One schema may express a similar "world view" with different syntax, grammar or structure
- One schema may be a new version of the other
- Two or more schemas may be evolutions of the same original schema
- There may be many sources modeling the same aspects of the underlying domain ("horizontal resolution" such as for competing trade associations or standards bodies), or
- There may be many sources that cover different domains but overlap at the seams ("vertical resolution" such as between pharmaceuticals and basic medicine).

Regardless, the needs for semantic mediation are manifest, as are the ways in which semantic heterogeneities may arise.

Classification of Semantic Heterogeneities

The first known classification scheme applied to data semantics that I am aware of is from William Kent nearly 20 years ago.[5] (If you know of earlier ones, please send me a note.) Kent's approach dealt more with structural mapping issues (see below) than differences in meaning, which he pointed to data dictionaries as potentially solving.

The most comprehensive schema I have yet encountered is from Pluempitiwiriyawej and Hammer, "A Classification Scheme for Semantic and Schematic Heterogeneities in XML Data Sources." [6] They classify heterogeneities into three broad classes:

- *Structural* conflicts arise when the schema of the sources representing related or overlapping data exhibit discrepancies. Structural conflicts can be detected when comparing the underlying DTDs. The class of structural conflicts includes generalization conflicts, aggregation conflicts, internal path discrepancy, missing items, element ordering, constraint and type mismatch, and naming conflicts between the element types and attribute names.
- *Domain* conflicts arise when the semantic of the data sources that will be integrated exhibit discrepancies. Domain conflicts can be detected by looking at the information contained in the DTDs and using knowledge about the underlying data domains. The class of domain conflicts includes schematic discrepancy, scale or unit, precision, and data representation conflicts.
- *Data* conflicts refer to discrepancies among similar or related data values across multiple sources. Data conflicts can only be detected by comparing the underlying DOCs. The class of data conflicts includes ID-value, missing data, incorrect spelling, and naming conflicts between the element contents and the attribute values.

Moreover, mismatches or conflicts can occur between set elements (a "population" mismatch) or attributes (a "description" mismatch).

The table below builds on Pluempitiwiriyawej and Hammer's schema by adding the fourth major explicit category of language, leading to about 40 distinct potential sources of semantic heterogeneities:

Class	Category	Subcategory
STRUCTURAL	Naming	Case Sensitivity
		Synonyms
		Acronyms
		Homonyms
	Generalization / Specialization	
	Aggregation	Intra-aggregation
		Inter-aggregation
	Internal Path Discrepancy	
	Missing Item	Content Discrepancy
		Attribute List Discrepancy
		Missing Attribute
		Missing Content

	Element Ordering	
	Constraint Mismatch	
	Type Mismatch	
DOMAIN	SchematicDiscrepancy	Element-value to Element-label Mapping
		Attribute-value to Element-label Mapping
		Element-value to Attribute-label Mapping
		Attribute-value to Attribute-label Mapping
	Scale or Units	
	Precision	
	DataRepresentation	Primitive Data Type
		Data Format
DATA	Naming	Case Sensitivity
		Synonyms
		Acronyms
		Homonyms
	ID Mismatch or Missing ID	
	Missing Data	
	Incorrect Spelling	
LANGUAGE	Encoding	Ingest Encoding Mismatch
		Ingest Encoding Lacking
		Query Encoding Mismatch
		Query Encoding Lacking
	Languages	Script Mismatches
		Parsing / Morphological Analysis Errors (many)
		Syntactical Errors (many)
		Semantic Errors (many)

Most of these line items are self-explanatory, but a few may not be:

- *Homonyms* refer to the same name referring to more than one concept, such as Name referring to a person v. Name referring to a book
- A *generalization/specialization* mismatch can occur when single items in one schema are related to multiple items in another schema, or vice versa. For example, one schema may refer to "phone" but the other schema has multiple elements such as "home phone," "work phone" and "cell phone"
- *Intra-aggregation* mismatches come when the same population is divided differently (Census *v.* Federal regions for states, or full person names *v.* first-middle-last, for examples) by schema, whereas *inter-aggregation* mismatches can come from sums or counts as added values
- Internal path discrepancies can arise from different source-target retrieval paths in two different schemas (for example, hierarchical structures where the elements are different levels of remove)
- The four sub-types of *schematic discrepancy* refer to where attribute and element names may be interchanged between schemas
- Under languages, *encoding* mismatches can occur when either the import or export of data to XML assumes the wrong encoding type. While XML is based on Unicode, it is important that source retrievals and issued queries be in the proper encoding of the source. For Web retrievals

this is very important, because only about 4% of all documents are in Unicode, and earlier BrightPlanet provided estimates there may be on the order of 25,000 language-encoding pairs presently on the Internet

- Even should the correct encoding be detected, there are significant differences in different language sources in *parsing* (white space, for example), *syntax* and *semantics* that can also lead to many error types.

It should be noted that a different take on classifying semantics and integration approaches is taken by Sheth et al.[7] Under their concept, they split semantics into three forms: implicit, formal and powerful. Implicit semantics are what is either largely present or can easily be extracted; formal languages, though relatively scarce, occur in the form of ontologies or other descriptive logics; and powerful (soft) semantics are fuzzy and not limited to rigid set-based assignments. Sheth et al.'s main point is that first-order logic (FOL) or descriptive logic is inadequate alone to properly capture the needed semantics.

From my viewpoint, Pluempitiwiriyawej and Hammer's [6] classification better lends itself to pragmatic tools and approaches, though the Sheth et al. approach also helps indicate what can be processed *in situ* from input data *v.* inferred or probabalistic matches.

Importance of Reference Standards

An attractive and compelling vision -- perhaps even a likely one -- is that standard reference ontologies become increasingly prevalent as time moves on and semantic mediation is seen as more of a mainstream problem. Certainly, a start on this has been seen with the use of the Dublin Core metadata initiative, and increasingly other associations, organizations, and major buyers are busy developing "standardized" or reference ontologies.[8] Indeed, there are now more than 10,000 ontologies available on the Web.[9] Insofar as these gain acceptance, semantic mediation can become an effort mostly at the periphery and not the core.

But, such is not the case today. Standards only have limited success and in targeted domains where incentives are strong. That acceptance and benefit threshold has yet to be reached on the Web. Until such time, a multiplicity of automated methods, semi-automated methods and gazetteers will all be required to help resolve these potential heterogeneities.

NOTE: This posting is part of an occasional series looking at a new category that I and BrightPlanet are terming the **eXtensible Semantic Data Model** (XSDM). Topics in this series cover *all information related to extensible data models and engines applicable to documents, metadata, attributes, semi-structured data, or the processing, storing and indexing of XML, RDF, OWL, or SKOS data. A major white paper will be produced at the conclusion of the series.*

[1] Michael Stonebraker and Joey Hellerstein, "What Goes Around Comes Around," in Joseph M. Hellerstein and Michael Stonebraker, editors, *Readings in Database Systems, Fourth Edition*, pp. 2-41, The MIT Press, Cambridge, MA, 2005. See http://mitpress.mit.edu/books/chapters/0262693143chapm1.pdf.[2] Natalya Noy, "Order from Chaos,"

ACM Queue vol. 3, no. 8, October 2005 See http://www.acmqueue.com/modules.php?name=Content&pa=showpage&pid=341&page=1

[3] Alon Halevy, "Why Your Data Won't Mix," *ACM Queue* vol. 3, no. 8, October 2005. See http://www.acmqueue.org/modules.php?name=Content&pa=showpage&pid=336.

[4] Michael K. Bergman, "The Deep Web: Surfacing Hidden Value," *BrightPlanet Corporation White Paper*, June 2000. The most recent version of the study was published by the University of Michigan's *Journal of Electronic Publishing* in July 2001. See http://www.press.umich.edu/jep/07-01/bergman.html.

[5] William Kent, "The Many Forms of a Single Fact", *Proceedings of the IEEE COMPCON*, Feb. 27-Mar. 3, 1989, San Francisco. Also HPL-SAL-88-8, Hewlett-Packard Laboratories, Oct. 21, 1988. [13 pp]. See http://www.bkent.net/Doc/manyform.htm.

[6] Charnyote Pluempitiwiriyawej and Joachim Hammer, "A Classification Scheme for Semantic and Schematic Heterogeneities in XML Data Sources," *Technical Report TR00-004*, University of Florida, Gainesville, FL, 36 pp., September 2000. See ftp.dbcenter.cise.ufl.edu/Pub/publications/tr00-004.pdf.

[7] Amit Sheth, Cartic Ramakrishnan and Christopher Thomas, "Semantics for the Semantic Web: The Implicit, the Formal and the Powerful," in *Int'l Journal on Semantic Web & Information Systems*, 1(1), 1-18, Jan-March 2005. See http://www.informatik.uni-trier.de/~ley/db/journals/ijswis/ijswis1.html

[8] See, among scores of possible examples, the NIEM (National Information Exchange Model) agreed to between the US Departments of Justice and Homeland Security; see http://www.niem.gov/.

[9] OWL Ontologies: When Machine Readable is Not Good Enough