Data Semantics: *what, where and how?*

A. Sheth
Large Scale Distributed Information Systems Lab,
Department of Computer Science, University of Georgia
415 GSRC, Athens, GA 30602-7404, USA
e-mail: amit@cs.uga.edu; URL:http://www.cs.uga.edu/LSDIS/

Abstract

At the panel held during the last session of the DS-6 conference, four panelists -- Leo Mark, Robert Meersman, Sham Navathe, and Arnon Rosenthal -- addressed the key questions related to the topic of the conference, related their perspectives to what was presented and discussed at the conference, and suggested research issues in data semantics that they would like to see addressed in the future. The panel was organized, introduced, and moderated by the author. Several conference participants also presented short position statements during the panel. This chapter summarizes the lively and often insightful panel discussion, along with additional thoughts of the author/moderator.

Keywords

Data Semantics, Application Semantics

1   INTRODUCTION

The panelists participating in the panel on "Data Semantics: what, where and how?" were asked to address four fundamental questions directly related to the subjects of this conference:
- What is data semantics?
- Where do you find it (what do you look for and where do you look to understand the semantics of data) and how do you derive it?
- How do you represent it?
- What are the uses of semantics (how do humans, applications, and data management systems use data semantics)?

   Several papers at the conference also dealt, either directly or indirectly, with the above questions. In addition to giving their own insights, the panelists pointed out and interpreted the papers presented at the conferences that dealt with these questions. They also offered their view on future research that could better address the questions posed to them. In writing this chapter that describes and summarizes the discussions, we have also

taken the liberty of adding a few additional comments and pointing to several relevant references to enrich the discussion.

   Section 2 reviews the discussions in response to the first question -- "what is data semantics?"   Section 3 addresses the second question -- "where do you get the data semantics and how do you derive it?" Section 4 presents the views on "how are (can) semantics (be) represented?" Section 5 addresses the question related to the uses of data semantics. Finally, Section 6 titled "Parting Thoughts" presents additional discussion of the topics addressed by the panel. In the following exposition, regular parentheses are usually used to further clarify a panelist's comment or observation, add to it, or provide our interpretation to it. Earlier collections of papers dealing with data semantics include Sheth (1991) and Hsiao et al. (1993).

2   THE BIG QUESTION: WHAT IS SEMANTICS?

Several perspectives on "What is Semantics?" were offered at the panel. During the panel introduction, we referred to the classic paper by Wood (1985), which defines semantics as the "*meaning* and the *use* of data". Still, a definition like this leaves a lot out -- what is a "meaning" and what is "use", and when do you know if you have adequately captured these to say that you have understood the semantics of that data?

   In the information systems context, semantics can be viewed as a mapping between an object modeled, represented and/or stored in an information system (e.g., an "object" in a database) and the real-world object(s) it represents. This mapping represents the semantics of the modeled object by describing or identifying the meaning and the use perspectives. Several DS-6 participants and panelists seemed to subscribe to this view.

   Rosenthal defined data semantics as a connection from a database to the real-world outside the database. He also considered the regularities in databases (i.e. constraints) that capture the regularities in the real-world (behavior?) as a component of the semantics. Mark and Meersman had similar perspectives on what semantics is.   Mark defined semantics as the "meaning of data" and "a reflection of the real world". Meersman defined semantics to be "a (set of) mapping(s) from your representation language to agreed concepts (objects, relationships, behavior) in the real-world".

   Mark sought to distinguish between "data semantics" and "data application semantics" and suggested that it is the latter that matters. This observation we think is interesting and relates well to Wood's view in the sense that applications provide the context of the use of the data.

   Meersman's presentation of his perspectives on the question was very interesting. The first slide he presented had one thing (object) on it - a (small) red dot. He asked the audience, "what is this?" (the question can also be taken as "what does this thing on the slide mean to you?"). In this process (according to our interpretation), he asked the audience to engage in a (thought) process of "determining the semantics" of the object that appeared as "a red dot"[*]. Meersman's exposition to the question "what is data semantics?" was inter-twined with the first half of another question, "how do you derive and represent it?" Specifically, he presented two beliefs (which he called his thesis):

---

[*] See Meersman (1994) for a more detailed exposition.

- there is no semantics without some form of (formal? informal?) agreement (between the agents observing the real-world), and
- semantics always exists but we need an *interpretation agent* to determine/derive the "meaning" (associated with an object), and consequently any semantics implies the existence of an agent (interpreter) interacting with a domain.

Based on David Beech's excellent keynote, Navathe defined semantics in broader terms as "Anything about data that has a conceivable practical consequence (in applications)". Two interesting points the he identified were:

- semantics is all pervasive and covers many things such as the interpretation and the use of data, or the interaction of people to convert data into information (that is, semantics is everywhere and has broad interpretation), and
- semantics is dependent on humans, thus it is difficult to address it in the context of machines (no wonder the systems oriented database researchers often find it a "soft science").

The three keynote speakers during the conferences also addressed some of the important issues related to semantics, which we would like to point to (these were also recalled during the panel). David Beech viewed semantics primarily in terms of capturing *similarities* between objects. Gio Wiederhold identified *relationships* between objects as the key to the semantics; this is indeed the perspective taken by many in the knowledge representation field and many in the audience seem to subscribe to it. However, defining the degrees of similarities or types of interesting relationships between objects is a hard problem and is briefly addressed in the section on representation of semantics. Jim Foley gave an excellent discussion of *context* as it related to World-Wide Web (WWW) and data visualization. His examples also showed the need to go beyond syntactic components.

3   WHERE DO YOU GET THE DATA SEMANTICS? HOW DO YOU DERIVE IT?

The essence of this question was what do you[*] do, where do you look at, and what do you analyze to understand the semantics, before the semantics can be represented in some notation, syntax and structure. For example, if you take Wood's definition of semantics, the question is how do you go about understanding the "meaning" and the "use" of data.

According to Meersman, semantics is derived from an agreement between cognitive agents observing the real-world. For example, (we interpret his proposal as) one looks at an agreement such as "this is a red dot" to further derive semantics, by asking what this means to the user/application/observer or how it is used. The complex issue this point presents to us is that the agreement itself (that this is a "red dot") is often not the semantics itself, but "something" that lends to determining what is semantics. Furthermore, what this "something" is, is itself the heart of the question (and is often poorly answered).

Meersman further described the process of deriving semantics as follows. He proposed that semantic concepts are presently based on reductionism. Interested parties (cognitive agents?) have to agree on a common observation, apply cognitive processes, determine if

---

[*] .You. stands for either a human or a program/system.

the pieces of the world overlap, and reverse engineer the complex cognitive processes to derive semantics from the database. He further commented that reductionism helps, but that you can only guess at the analytical processes involved. Semantics always exists, but we need interpretation agents for deriving the meaning (i.e., an existence of "outside" agents is needed) and that it is useless to talk about semantics without an agreement.

Towards the end of the panel, Meersman had additional observations relevant to this question. He noted that methods and processes have an important contribution to data semantics (we view this comment as "look at the methods and processes that use or operate on data to derive data semantics"). He identified two components related to methods and processes: "what they do" (i.e., functional component) which could also be determined from requirements or interrogating users, and "how they do it" (i.e., procedural component).

Navathe had several observations on the types of semantics and where to look for semantics. He noted that

- semantics lie in the way people interact with (access and update?), communicate, interpret, and use data or the information implicitly presented in the data (we sense a slightly circular argument in the second phrase, since "information" means the ability to understand data (i.e., data semantics) and use the data for decision making),
- semantics (termed "software semantics") can be found from data structure or knowledge representation schemas, programs, communication protocols, and encoding,
- semantics can be observed/found in hard-copy form in books, manuals, policy guides, and legal documents (perhaps a beneficial way to see this point, for example, would be the use of a policy guide to interpret data to obtain semantics; we would find it hard to agree a print form itself as the semantics), and
- semantics (termed "procedural semantics") can be observed in or obtained from programmed procedures, social protocols, and law.

Saying that "we do not need machines that simulate hand-shakes", Navathe strongly advocated the need to involve humans, not just machines, in all aspects of dealing with semantics. He also felt that, new models that better capture extracted knowledge representing the semantics, need to be developed.

The rest of the panelists had a somewhat more practical view of the question being discussed in this section. Mark noted that one can get the semantics (of data) "from the real-world applications (that use the data)". Mark also noted that one of the associated problems is that the types of applications have exploded (consequently making it harder to extract semantics based on every type of application). We mainly see Mark's response to this question as the "use" perspective in Wood's definition. We still think that the "meaning" perspective is also important. One reason for this is that existing real-world applications may not be fully exploiting the uses of the data. More uses can be facilitated if the semantics were to capture what the data means, whether in the context of current applications/uses or other future ones.

On a related note, giving the example of deriving the semantics to make improvements to the database management systems, Mark suggested looking at data carefully and tracking the execution of programs. Specific instances which he pointed out included (a) an index which keeps track of how often queries are run and which data are used in the

database system, (b) a "cacher" which keeps tracks of the queries and creates an index to point to the exact point where the data corresponding to the meaning is stored, (c) rules that try the execution of transactions and refine the concurrency control programs together, and (d) semantics of operations to allow normally conflicting operations to execute. On the question of how to derive semantics, Mark identified three players: users, domain experts, and (programmatic evaluations such as) data mining.

Rosenthal had proposed looking for the regularities in the databases to look for semantics. The paper by Srinivasan et al. (1995)[*] uses a clustering technique including data usage patterns on data to unearth semantic similarity between data objects managed in different databases.

4   ON REPRESENTING SEMANTICS

To present our views on the topic of representing semantics, we referred to the talk at DS-5 (Sheth and Kashyap(1992)) where we had introduced the concept of *semantic proximity* to represent semantics. In hindsight, we would like to note that it captures (or at least attempts to capture) the views on semantics presented by all three keynote speakers (see the last paragraph in Section 2). Semantic proximity captures degrees of similarities or types of relationships between the (model world) objects, and uses context as a key component of the representation. While *context* has been recognized as a key component of semantics (and we view it as something that underpins the semantics of an object represented in the model world), it was not adequately addressed by the panel. There are many widely differing notions of contexts (e.g., for extended discussion, see Kashyap and Sheth (1995)) including one presented by Michael Siegel (Daruwala et al. (1995)) as well as in the keynote by Foley as mentioned in Section 2.

Navathe identified several modeling abstractions or mechanisms that can be used to represent semantics and pointed out the papers presented at this conference that used them. These included Context by Daruwala et al. (1995), Multidatabase manipulation language by Misser and Rusinkiewicz (1995), Formal language by Weigand et al. (1995), a combination of language, methods and tools by van Keulen et al. (1995), Rules by Herbst et al. (1995), Roles by Wong and Li (1995), Views by Al-Anzi and Spooner (1995), Semantic integrity constraints by Embury and Gray (1995), Cases by Zeleznikow et al. (1995), Default values by Halpin and Vermeier (1995), and Menus by Foley (1995). It might be accurate to say that there were as many semantic representation formalisms and methods as the number of researchers at the conference.

Rosenthal emphasized the need for methodologies to be expressed in (or translated to) simple terms, preferably monosyllables, that pragmatists can understand. We researchers sometimes do not check enough to see if the questions in our methodologies are answerable. For example, a question like "do these types mean the same thing?" may not be meaningful. One needs a connection to more concrete things, e.g., "do these two types describe the same set of objects in the external world" or (when describing conceptual

---

[*] All citations such as this, which are not given at the end of this chapter for brevity, appear as chapters in this book.

schemas and ontologies) "if applied to the same organization, would these two types have the same set of instances?" (Objects organize information about the external world; types are higher order abstractions that organize objects).

As an illustration of how formalisms can cause people to turn off common sense, Rosenthal mentioned that engineers at NASA did not protest when their methodology on fault tree analysis gave answers for probability of failure that was three orders of magnitude less than any of the engineer's intuitions. He asked "perhaps we methodology developers sometimes fall prey to the same blindness?"

On the matter of representation of semantics, Rosenthal proposed that (a) one should consider the issues of reuse, evolution, and merging (with the tool that may use the data), and (b) modularity is crucial for meta-database. We interpret these to be his criteria for representation of semantics (i.e., good semantic representation should support reuse, evolution, and merging of the data). On the matter of deriving semantics, he identified human insight as the key ingredient, followed by the use of tools (e.g., information retrieval tool looking for similarities between text) supported by human confirmation of the findings.

Later in his talk, Rosenthal sought to distinguish between "lite" semantics and "heavy" semantics. He defined the former to be clear and deterministic, containing simple information or information in small "chunks", involving simple (representation) formalism, and involving simple mediation services. Rosenthal used property-value list, glossary, and "fancy formalisms/AI" to exemplify what he meant by heavy semantics. He suggested that the techniques used to discover semantics determine whether one obtains light or heavy semantics. At times we have used terms "weak semantics" to mean the semantics that can be identified based on structural, syntactic, and value/extensional information in databases. We have used the term "deep semantics" when dealing with semantics involves the issues of human cognition, perception, or interpretation. The papers in the conferences primarily dealt with weak semantics, which is consistent with the general realization that it is very hard to deal with the latter.

Relationships between objects as the key to semantics was mentioned earlier with reference to Wiederhold's talk. The paper by Waesch and Aberer (1995) discusses modeling relationships. Most of the efforts in this area, however, have focused on modeling structural relationships (and hence deal with "weak semantics"). Readers can find a good discussion of a variety of relationships in Hammer and McLeod (1993). In this context, we also wish to point out Kent's classical paper titled "Many Forms of the Same Fact" (Kent (1989)).

On the topic of representing semantics, Mark presented a hierarchy consisting of adequate models (that include dynamic and time aspects), structural constraints and behavioral constraints. He also identified the tension between declarative *versus* procedural paradigms for presentations. Further addressing the question of representation of semantics, Mark said it is a *déjà vu* all over again. He referred to the discussion about data models involving the need for declarative specifications in the mid-70s, advocacy of O-O data models in the 80s, and subsequent ceremony to bury data models at the meeting of database researchers at Laguna Beach (Neuhold and Stonebraker, 1988). Nevertheless the view of many is that the latter proclamation of the elite did not serve as a marching order on the troops, and the issues related to data structures and modeling are still alive

and kicking. Mark suggested that we are doing it all over again with respect to WWW (and we may add, also with respect to metadata for heterogeneous digital data). Foley's insightful exposition as well as panel discussions at many conferences and workshops have identified the limitations of the current representation formalisms of modeling WWW data, and have pointed to the need for enriching, modifying or replacing the representation formalisms/models. Getting more out of data while keeping the representation models "lite" will be a challenge we will face all over again.

5 ON THE USE OF SEMANTICS

The last question the panelists faced was "what are the uses of semantics, and how is semantics used? " We had framed the second part ("how is semantics used? ") in terms of the possible use of semantics in query processing and optimization. This question received little attention and time from the panelists, perhaps because of time limitations.

Navathe identified several examples of how semantics was used in some of the papers presented at the conference. Misser and Rusinkiewicz (1995) used and addressed semantic issues to support multidatabase manipulation, Rosenthal and Sciore (1995) used semantics to achieve better semantic (in the sense of "more meaningful") interoperability in a distributed object management environment, Atzeni and Torlone (1995) used it to support schema/model translation, Yoon and Kerschberg (1995) used it to support update propagation while satisfying integrity constraints, and Comai et al. (1995) capture and use semantics in active database systems to under-pin the execution behavior of rules.

6 PARTING THOUGHTS

Many database researchers (in higher percentage at this meeting than many other database conferences) now better appreciate the limitation of purely syntactical approaches in dealing with data. The syntactic approaches do not lead us too far in obtaining information from the data (or in using information to create better, more useful data). At the same time, progress is slow in dealing with semantics since it is something that cannot be captured completely, cannot be done programatically (alone) or fully automated, does not seem to have a purely mathematical or formal model, and requires humans to participate with as much support from the computer systems as possible.

In this print media, primarily using text, it is impossible to capture all the dynamics and nuances of a lively discussion and debate. A multimedia form of this exposition could improve the situation a bit but would still not be able to accurately duplicate the panel. The same is true for a model world in which we can provide the syntax and representation of a real-world object, then improve the situation somewhat by capturing some of the "weak" semantics using structural relationships, enrich it further with additional knowledge, and hopefully capture the interesting and important aspects of the real-world object represented in an information system.

We all know that we are far from adequately capturing the semantics that could be derived using all the senses, cognition, perceptions, and interactions between multiple

agents. Nevertheless, the gain of capturing even the limited amount of useful semantics can be tremendous if we want the "system" to be more "intelligent" or to have the "system" assist the user in a more "intelligent" way. In the discussion following Navathe's talk, David Beech shared an interesting thought. He noted that just as Newton's laws were refined by Einstein's laws (that is, the scientific progress marked by the latter did not invalidate the former), he expects the progress in (research issues involving) semantics to be similar. In the introduction to the panel, we discussed that the real world consists of and has to deal with inconsistency, incompleteness and uncertainty; these aspects need to be addressed with respect to the questions on semantics posed for this panel.

Mark and several other panelists brought up the issues of the World-Wide Web (WWW) as a very large database with its "spaghetti" structure and its failings as an information-base due to its lack of support for semantic issues. This topic was discussed in more detail during the panel organized by Erich Neuhold at the conference, and we refer the reader to the material based on that panel, as well as Berwick et al. (1995).

After the presentations by the panelists and some discussion, we had invited several members from the audience to share their thoughts in a rather short amount of time. We would like to end this re-cap of the panel with the very last of these presentations, which also seemed to provide fodder for more than the usual share of discussions that continued after the final session. Vipul Kashyap made his points using the following figure. We think the figure is largely self descriptive, and hence we will not describe it further (more details are available at URL: http://www.cs.uga.edu/LSDIS/).
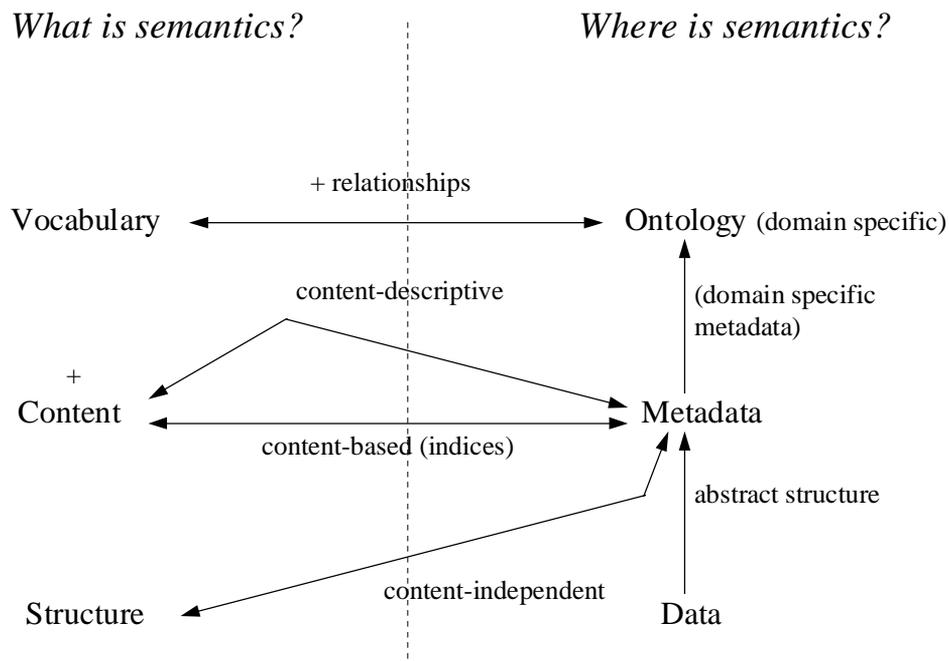

Figure 1 Semantics-- What and Where?

7 ACKNOWLEDGMENT

8 REFERENCES

Berwick R., Carrol J., Connolly C., Foley J., Fox E., Imielinski T., and Subramanian V. (1995) Research Priorities for the World-Wide Web. Report of the NSF Sponsored Workshop, J. Foley and J. Pitkow, Eds., April. Available at URL:*http://www.cc.gatech.edu/gvu/nsf-ws/report/Report.html*.

Hammer, J. and McLeod, D. (1993) An Approach to Resolving Semantic Heterogeneity in a Federation of Autonomous Heterogeneous Database Systems, *Intl. Journal of Intelligent and Cooperative Information Systems*, 2 (1).

Hsiao D., Neuhold E. and Sacks-Davis R. (eds.) (1993) *Interoperable Database Systems* (*Proc. of the DS-5 Conference on Semantics of Interoperable Database Systems*, November 1992, Lorne, Australia), IFIP Transactions A-25, North-Holland.

Kashyap, V. and Sheth, A. (1995) Schematic and Semantic Similarities between Database Objects: A Context-based Approach, Technical Report TR-CS-95-001, LSDIS Lab, Dept. of Computer Science, University of Georgia, January. [Available at: URL: *http://www.cs.uga.edu/LSDIS/pub_METADATA.html*; a shorter version will appear in the VLDB Journal.]

Kent, W. (1989) The Many Forms of a Single Fact, *Proc. IEEE COMPCON*, San Francisco, CA.

Meersman, R. (1994) Some Methodology and Representation Problems for the Semantics of Prosaic Application Domains, *Proc. of the ISMIS'94 Conference*, Z. Ras and M. Zemankova (eds.), Springer Verlag.

Neuhold, E. and Stonebraker, M. (1988) Future Directions in Database Research, TR-88-001, ICSI Report.

Sheth, A. (ed.) (1991) Special Issue of SIGMOD Record on `Semantic Issues in Multidatabase Systems`, 20 (4), December.

Sheth, A. and Kashyap, V. (1993) So Far (Schematically) yet So Near (Semantically), (invited paper based on a keynote talk), in Hsiao et al. (1993).

Wood, J. (1985) What's in a link? *Readings in Knowledge Representation*, Morgan Kaufmann.

The following citations in this chapter appear in this book:
Al-Anzi and Spooner (1995); Atzeni and Torlone (1995); D. Beech (1995); Comai et al. (1995).; Daruwala et al. (1995); Embury and Gray (1995); Foley (1995); Halpin and Vermeier (1995); Herbst et al. (1995); Misser and Rusinkiewicz (1995); Rosenthal and Sciore (1995); Srinivasan et al. (1995), van Keulen et al. (1995); Waesch and Aberer (1995); Wiederhold (1995); Wong and Li (1995); Yoon and Kerschberg (1995); Zeleznikow et al. (1995).