Chapter 2.35
Data Semantics

Daniel W. Gillman
Bureau of Labor Statistics, USA

INTRODUCTION

Almost every organization, public or private, for profit or non-profit, manages data in some way. Data is a major corporate resource. It is produced, analyzed, stored, and disseminated. And, it is poorly documented.

Descriptions of data are essential for their proper understanding and use by people inside and outside the organization. For instance, systems for disseminating data on the Internet require these descriptions (Census Bureau, n.d.). Either inside or outside the organization, functions of the system support finding the right data for a study, understanding data from a particular source, and comparing data across sources or time (Gillman, Appel, & LaPlant, 1996).

Descriptions of data and other resources are metadata (Gillman, 2003). Metadata are part of the corporate memory for the organization, and preserving corporate memory is one of the basic features of knowledge management (King, Marks, & McCoy, 2002). Metadata include the meaning, or semantics, of the data. In some countries, such as the U.S., a large percentage of the population is reaching retirement age. As a result, recording the memories of these workers, including the meaning of data, is increasingly important. Preserving metadata is crucial for understanding data years after the data were created (Gillman et al., 1996).

Traditionally, the metadata for databases and files is developed individually, without reference to similar data in other sources. Even when metadata exist, they are often incomplete or incompatible across systems. As a result, the semantics of the data contained in these databases and files are poorly understood. In addition, the metadata often disappear after the data reach the end of the business lifecycle.

Techniques for documenting data are varied. There are CASE (Computer-Aided Software Engineering) tools such as Oracle Designer® (Oracle, n.d.) or Rational Rose® (IBM, n.d.). These tools produce models of data in databases (Ullman, 1982). The models provide some semantics for the

data. For social science data sets, metadata is described in an XML (eXtensible Markup Language) specification (ICPSR, n.d.). For geographic data sets, the U.S. Federal Geographic Data Committee developed a metadata framework, clearinghouse, and supporting software (FGDC, n.d.).

Metadata are data, too. They are structured, semi-structured, or unstructured (Abiteboul, Buneman, & Suciu, 2000), just as data are. Data are structured if one knows both the schema and datatype, semi-structured if one knows one of them, and unstructured otherwise. From the perspective of their content, documents are unstructured or semi-structured data. Their schemas come from presentation frameworks such as HTML (Hyper-Text Mark-up Language) (W3C, 1997) or word processor formats. Documents with the content marked up in XML (W3C, 2004) are semi-structured. When using the full datatyping capability of XML-Schema, the document is structured with respect to the content. However, the colloquial use of the term "document" begins to lose its meaning here.

In describing some resource, the content is more important than the presentation. The content contains the semantics associated with the resource. If the content is structured data, this increases the capability of performing complex queries on it. Retrieving unstructured documents using search engine technology is not as precise.

It turns out there are structured ways to represent the semantics of data. Ontologies (Sowa, 2000) are the newest technique. Traditional database (or registry) models are examples of ontologies. This article describes the constituents of the semantics of data and a technique to manage them using a metadata registry. The process of registration—an approach to control the identification, provenance, and quality of the content—is also described and its benefits discussed.

SEMANTICS OF DATA

Terminology

To begin, we describe some useful constructs from the theory of terminology. These come from several sources (Sager, 1990; ISO, 1999, 2000). We use these constructs to describe the semantics of data. The terms and definitions follow in a list below:

- Characteristic: Abstraction of a property of a set of objects.
- Concept: Mental constructs, units of thought, or unit of knowledge created by a unique combination of characteristics.
- Concept system: Set of concepts structured according to the relations among them.
- Definition: Expression of a concept through natural language, which specifies a unique intension and extension.
- Designation: Representation of a concept by a sign, which denotes it.
- Extension: Set of objects to which a concept refers.
- General concept: Concept with two or more objects that correspond to it (e.g., planet, tower).
- Generic concept: Concept in a generic relation having the narrower intension.
- Generic relation: Relation between two concepts where the intension of one of the concepts includes that of the other concept and at least one additional distinguishing characteristic.
- Individual concept: Concept with one object that corresponds to it (e.g., Saturn, Eiffel Tower).
- Intension: Sum of characteristics that constitute a concept.
- Object: Something conceivable or perceivable.