

Knowledge and Dataspaces

Basically, our goal is to organize the world's information and to make it universally accessible and useful.—Larry Page, Google¹

What's in This Section?

Google gains power from knowledge about information, users, and system processes. The Programmable Search Engine allows the company to apply a specific method to building knowledgebases. However, conflicts exist among data. Google has invested in technology that can resolve such conflicts. The combination of the two Google methods make it possible for the company to generate new, high-value reports, compilations, data, and other information objects.

Knowledge Is Power

It is a truism to be sure, but the phrase “knowledge is power” contains a useful message. Few people take the time to look at Google's technical papers. Even fewer pay attention to Google's patent documents. Yet these two sources provide an observation deck from which one can view some of the inner workings of the global, evolving Google infrastructure of smart software, smart hardware, and even smarter engineers and scientists.

Here is an example.

In February 2007, Google disclosed that it had a system based on smart software and other Google functions that could perform several useful functions - follow along by downloading and examining the technical diagrams in “Programmable Search Engine”, invented by Ramanathan V. Guha.²

The PSE is a system that imparts semantic metatags to content. Its inventor worked on the World Wide Web Consortium's semantic Web standard. The five patent applications that make up this invention reveal a system that captures meaning of documents and associated objects. The tags don't exist in a vacuum. A context is generated for each object. Together with usage data and other information in the Googleplex, the PSE has some interesting and potentially far-reaching implications.

A Web site operator can “push” data to Google so Google can index it. If the Web site operator does not want to schedule uploads of data to Google using the Google Base mechanism. Google can index the content on the public Web site. If the Web site generates pages dynamically or requires a user name and password. Google can handle those functions. If the Web site does not want Google to index its data, Google will respect the block expressed in the Web site's robots.txt file.³

Google's PSE makes it possible to ingest data, combine them, and create a master data repository automatically. The PSE straightens out problems and fills in missing fields. if a Web site focusing on consumer camera products, Google will merge the consumer data and metadata with information from

1. From http://www.woopidoo.com/business_quotes/authors/larry-page-quotes.htm

2. US2007/0038616, 33 pages. Cross references exist in US2007/0038601 and US2007/0038603.

3. A robots.txt file is convention designed to prevent cooperating Web spiders and other Web indexing mechanisms from accessing content which may otherwise be publicly available.

a Web site with information about cameras for a professional photographer. The idea is to:

- Normalize data that would normally be in different forms or have some important pieces of information missing such as an educational institution's price or special order options
- Organize the data so that it can be sliced and diced via software without involving a human in the process
- Associate the data with the Web site providing one or more items of data and the other camera-related information
- Update the individual items of data so that the index remains "fresh" or current. Prices change frequently and when a person looks for a product, certain items of information are "must have" facts like availability, price, etc.

From the user's point of view, when a product—for example, a camera—is the subject of a query, Google notes the product, what cameras were included in the results list, and what particular results or data generated the most clicks. Google performs this type of abstract usage tracking 24x7 and makes the results available to other Google functions. The Programmable Search Engine processes the log files in order to generate additional information about the set of camera data. For example, Google attempts to determine:

- What related data did the user access when reviewing camera data; for example, Google Maps?
- How long did the user spend researching cameras and how much "dwell" time was invested in a particular topic, camera, or Web site?
- What actions did the user take when reviewing the camera data; for example, browsing results to locate a Google Checkout vendor and placing an order, choosing delivery options, and so on?

The idea is to generate additional metadata to represent the information and the user behavior.

With the metadata and the structured information, Google has a unique knowledgebase, described in the Google open source using the term "context". The connotation of context is intended to get information about what information and data are needed and used in what situations and the machine processes involved to produce the outputs. In short, Google's notion of context is an umbrella capturing information about information, topics and concepts, users, and machine processes.

Google's Programmable Search Engine was been discussed in my 2007 monograph, *Google Version 2.0*.⁴ As impressive as the semantics and the knowledgebase are, some difficult issues remain. For example, what does Google's software do when the system identifies multiple addresses for a business? Does Google alphabetize the addresses? Does Google use some other log file data to identify the most important address? What about situations where the terminology is different but the product, service, or concept is the same? How can Google reconcile linguistic differences in one language? What about multi-lingual information; for example, camera manufacturer information in Japanese versus manufacturer information in German?

4. The monograph is available at <http://www.infonortics.com/publications/google/google-predator.html>

To achieve this knowledge representation, Dr. Halevy developed techniques that could deal with seemingly intractable problems of semantic heterogeneity, clashes in data itself such as different telephone numbers for an individual, and the meaning of ambiguous or incomplete information.

Since 1993, dataspace technology has moved in a stair step fashion. The core level is the technology required to make sense of disparate data types and models. Next was a series of demonstration projects conducted by researchers at the University of Washington and elsewhere (not summarized here) that provided insights into the algorithms and iterative procedures required to transform data. The Nimble technology was tangible evidence that automated normalization was sufficiently hardened for commercial use. With Transformic, Dr. Halevy applied his learnings to the far larger, more variable content available on the public Internet. If it so chooses, Google's acquisition of Transformic along with Dr. Halevy and Jayant Mandravan, connections with Stanford's Dr. Janet Widom and other specialists in the dataspace discipline equips it to apply the systems and methods to a number of Google products and services.

Dr. Halevy arrived at Google with technology refined and modified from 1992 to 2005. At Google he has a computing platform that is known to scale, contain proprietary systems and methods for processing petabytes of data, and be positioned to make use of Dr. Halevy's dataspace research.

An evolving and self-managing system able to integrated Web scale data, dataspace technology has some potential to add utility to search and retrieval.

How Does a Dataspace System Work?

A dataspace system must perform operations to reconcile differences in representations of information. Dataspaces uses a technique called "semantic mapping." The idea is that software can figure out how field "Company_Name" is the same as field "CoName". Instead of performing massive lookups as part of content processing, a dataspace system generates a representation. "Lookups" occur only when they are needed to answer a query. Dr. Halevy calls this "pay as you go" processing.

The knowledge representation developed by Dr. Halevy and his colleagues discerns new relationships among, between, and within processed content. For example, in a lecture given in 2006 "Principles of Dataspace Systems", he identified these index items:

- DerivedFrom
- SnapshotOf
- HighlyCorrelatedWith.

With this type of indexing, dataspace make possible new types of queries. For example, a dataspace can process queries about lineage and uncertainty. In effect, the system allows a user or a process to determine the provenance of information and calculate a score that indicates the likelihood that an item, fact, or other piece of information is likely to be correct.

In order to make use of a dataspace, a management system is required. Dr. Halevy calls this "a dataspace support platform" or DSSP.

What does a dataspace "look" like? First, a dataspace is an abstraction of the information. It contains information comparable to an index for any search system. The system also contains unique index

information about the data sources themselves. A dataspace can also make use of other information such as usage behavior of users, data about particular systems and subsystems, and metadata that go beyond the meaning of a particular fact.

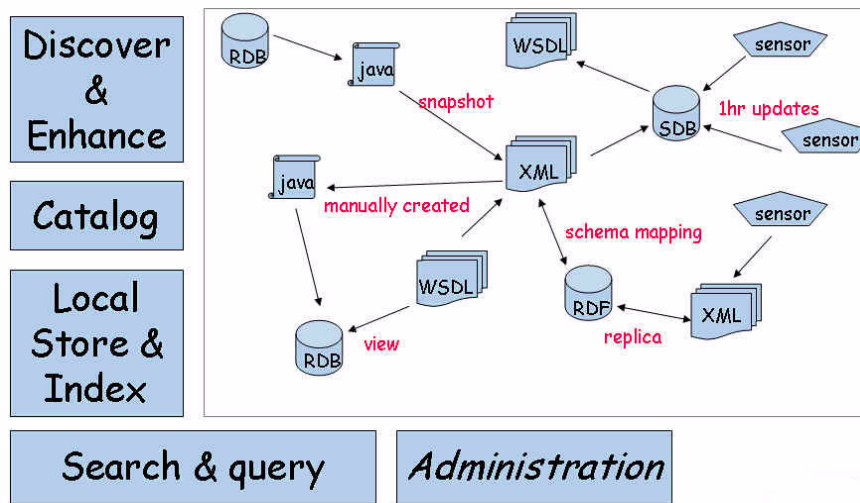
I want to look at one example of a representative type of metadata that a dataspace can capture. This illustration comes from one of Dr. Halevy's public lectures.

Several points warrant brief comment:

- Dr. Levy suggests that there are four core functions a dataspace and its support platform must perform. These are the ability to [a] process uncertain data, [b] update uncertainty values when new evidence becomes available to the system, [c] be proactive about reducing uncertainty when data clash, and [d] leverage lineage (provenance) to reduce uncertainty.
- The dataspace normalizes disparate types of content, making decisions about normalizing the data, modifying a master schema, and capturing information about participants and relationships
- The dataspace makes possible different types of queries and information analyses, described in the diagram as “discover and enhance” operations
- The lower-level operations required for transforming a limited range of content in the Nimble Technologies' system has been wrapped with more sophisticated administrative and operational processes; that is, the focus has shifted from performing transformations to building a system; that is, a dataspace support platform.

What can you do with a dataspace?

Google provides little information about applications of a dataspace. However, one of Dr. Halevy's colleagues—Dr. Jennifer Widom, a professor at Stanford University—provides a fascinating glimpse into dataspace's usefulness in law enforcement, insurance investigations, and business applications where some information is fuzzy, uncertain, or contradictory. In her discussion of the Trio system, a demonstration project using dataspace methods, Dr. Widom points out that “neither uncertainty nor lineage is supported in current database systems.”



This diagram shows the processing of disparate data. XML is the data type used in the system. The five rectangles outside the dataspaces process represent the components of the dataspaces support platform; that is, administrative tools for set up, search, indexing, cataloging operations, and the functions of discovery and enhancement of the knowledge representation in the dataspaces

This image is from Dr. Alon Halevy's presentation "Principles of Dataspace Systems," June 26, 2006. © Google, Inc. 2006

The example she presents involves a series of witnesses and their statements about a minor automobile accident. The information is processed and a series of mathematical recipes applied. The output is a report that generates truth tables with confidence scores generated by dataspaces methodology:

What Are the Implications of Dataspace?

Dataspace technology represents a significant shift in data management, data integration, and database systems. The challenges of petabyte data sets have hamstrung many organizations trying to make traditional work processes and relational databases work in a cost effective way.

Dataspace technology could change the rules for data management. Google could use its existing infrastructure to perform automatic transformation, advanced processes based on smart software, and introduce new types of queries to users.

If Google moves forward with its dataspaces innovations, there could be significant implications for traditional database and data management vendors. These companies do not have Google's ability to scale, nor its access to vast quantities of data to use to resolve certain ambiguities. Companies such as IBM, Microsoft, Oracle, and Sybase are aware of the limitations of traditional databases, and dataspaces technology is not new. Google seems uniquely positioned to use this technology to expand its system's utility.

My research suggests that Google could, without much fanfare, make its dataspaces technology available to law enforcement, financial services, and companies with large datasets and significant quantities of heterogeneous data. If Google company takes that action, it is in a position to add a significant feature to its existing services, particularly those tailored to an enterprise.

For consumer services, dataspaces technology would add interesting analytic capabilities to advertising. For the average Google Web search user, the system could provide needed information to generate reports, thus increasing its utility and possibly its share of the search market.

Applied to publishing, the dataspace technology appears to have a contribution to make for the dossier function (reconciling conflicting information about whereabouts and aliases), directories (resolving address and telephone number conflicts), determining values to determine reliability, accuracy, or relevance of facts or information objects. In short, the dataspace invention makes possible new types of auto-generated outputs because the system can deal with uncertainty and varied lineage of information. Dataspaces, in effect, can work like a publisher and a fact checker combined in one smart digital system.

An Example

Here is a sample output from Stanford University's Dr. Jennifer Widom, Professor of Computer Science and Electrical Engineering. Dr. Widom is a colleague of Google's Dr. Halevy. Here's a query output from one of Dr. Widom's lectures:

PrimeSuspect (crime#, suspect, accuser)	
1	Jimmy, Amy Billy, Betty Hank, Cathy
2	Frank, Cathy Freddy, Betty

Credibility (person,score)	
Amy	10
Betty	15
Cathy	5

List suspects with conf values based on accuser credibility

Suspects	
Jimmy 0.33 Billy 0.5 Hank 0.166	
Frank 0.25 Freddy 0.75	

The output of this dataspace's query is a list of suspects involved in a crime. The scores identify the most likely suspect. A police investigator would seek to talk with Freddy. © Dr. Jennifer Widom and Stanford University, 2008.

If Google commercializes such systems and methods as the Programmable Search Engine and dataspaces, it can sell reports such as:

- Special reports for intelligence (business and military) or law enforcement, litigation, and regulatory entities
- Rankings of investments or investment advisors
- Market analyses that include user, product, and system components; for example, uptake in a particular camera and most probable supporting products and services required by buyers of a model or class of camera

Dataspaces are a way to manage information and make it possible to run certain types of queries impractical in traditional indexes and databases. Dataspaces are constructs that integrate many separate indexes, their metatags, and data. In a dataspace, the user no longer worries about a specific collection or even what type of information has been processed-whatever the dataspace system has processed is available. Google's "universal search" is one of the first glimpses of the dataspace technology in the Google system you and I use for Web searching.