

Building Ontologies and Integrating Data from Multiple Agencies: A Case Study Using Gasoline

Jose-Luis Ambite* and Yigal Arens* and Luis Gravano† and Vasilis Hatzivassiloglou† and Eduard Hovy* and Judith Klavans† and Andrew Philpot* and Usha Ramachandran* and Jay Sandhaus† and Anurag Singla† and Brian Whitman†

*Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
hovy@isi.edu

†Department of Computer Science
Columbia University
535 West 114th Street, MC 1103
New York, NY 10027
klavans@cs.columbia.edu

Abstract

The massive amount of statistical and text data available from Federal Agencies has created a set of daunting challenges to both research and analysis communities. These problems include heterogeneity, size, distribution, and control of terminology. We are investigating solutions to three key problems, namely, (1) ontological mappings for terminology standardization; (2) data integration across data bases with high speed query processing; and (3) interfaces for query input and presentation of results. This collaboration between researchers from Columbia University and the Information Sciences Institute of the University of Southern California employs technology developed at both locations, in particular the SENSUS ontology, the SIMS multi-database access planner, the LKB automated dictionary and terminology analysis system, and others. The pilot application targets gasoline data from BLS, EIA, Census, and other agencies.

1 Introduction: The DGRC

As access to the web becomes a household commodity, the Government (and in particular Federal Agencies such as the Census Bureau, the Bureau of Labor Statistics, and others) has a mandate to make its information available to the public. But the massive amount of statistical and text data available from such agencies has created a set of daunting challenges to the research and analysis communities. These challenges stem from the heterogeneity, size, distribution, and disparity of terminology of the data. Equally, they stem from the need to provide broad and easy access to (and support proper understanding of) complex data.



DIGITAL GOVERNMENT RESEARCH CENTER

The Digital Government Research Center (DGRC) was established to address these problems. The DGRC consists of faculty, staff, and students at the Information Science Institute

(ISI) of the University of Southern California and Columbia University's Computer Science Department and its Center for Research on Information Access. The mandate of the DGRC is to conduct and support research in key areas of information systems, to develop standards/interfaces and infrastructure, build pilot systems, and collaborate closely with Government service/information providers and users.

Since its formation in 1999, the DGRC's activities include the EDC project and organizing and hosting the dg.o workshop in Los Angeles in May 2000 (see <http://www.dgrc.org>).

2 The EDC Project

2.1 Background: The Energy Data Collection

The DGRC Energy Data Collection (EDC) Project was started in the National Science Foundation's Digital Government program in 1999. It is developing solutions to three key problems in accessing large distributed data

collections: (1) sophisticated planning of access to multiple distributed and heterogeneous databases; (2) the use of a large ontology for terminology standardization and user guidance; (3) flexible user interfaces for query input, ontology browsing, and result presentation. A review paper is (Ambite et al., 2000).

In this work, the proposing team is working with representatives of major Federal and State statistics agencies and other organizations and individuals on a regular basis, to collect and disseminate statistical data. Representatives of these agencies, primarily from the Census Bureau, the Bureau of Labor Statistics (BLS), and the Energy Information Administration (EIA) of the Department of Energy (DoE), and the California Energy Commission (CEC). Other agencies we have met with include the National Center for Health Statistics (NCHS) and the Los Angeles County Administration. The Energy Information Administration provides extensive *monthly energy data* to the public on its Internet site (<http://www.eia.doe.gov>). The site is heavily browsed, receiving hundreds of thousands of hits a month, even though most of the information is available only as downloads of standard web (HTML) pages or as prepared PDF documents. Monthly data can only be obtained for the last few years in this manner. Current facilities thus provide only limited access to this very rich data source. Some portion of the data is also accessible by querying, but there are two serious problems hampering the current query system. First, it does not provide visibility for the many definitions and footnotes that explain the complex nature of the data and to changes that occur in series over time. Lack of awareness of such explanatory information often makes incomparable figures appear to be comparable. Second, the difficulty of defining queries makes the querying system useful only to expert users.

The EDC Project is addressing both problems. Techniques are being developed and implemented to attempt to make the complexities of data series either transparent, or more visible to users, depending on whether they can be handled independently by the system or not. And novel query facilities and other data analysis and presentation capabilities are being developed, that will be usable by the more common potential user of EIA data—the non-expert. Our aim is to build a system that, though still a research prototype, will be of benefit to various segments of the public, judging by the high demand for monthly energy data. Besides the

large number of browsers accessing EIA’s Web site, hardcopy publications further disseminate EIA’s monthly energy data. Such paper publications include the Monthly Energy Review and the International Petroleum Statistics Report. We will make the monthly data available through EDC’s new query system on the EIA’s web site.

In the first year of research and development for the Energy Data Consortium, we have demonstrated initial results in three areas:

Information Integration. The research questions we have addressed include effective methods to identify and describe the contents of databases so that useful information can be accurately and efficiently located even when precise answers are unavailable. We have wrapped over 120 databases for testing the first stage of information integration, and completed research on computational properties of aggregation, a necessary step to implementation;

Ontology Construction. We have built on the large formal SENSUS ontology at USC/ISI currently containing over 70,000 terms, linked together into a subsumption (*isa*) network, with additional links for *part-of*, *pertains-to*, and so on. Our results include extending the SENSUS high-level general domain ontology to incorporate new terms, extending and developing new automated concept-to-ontology alignment algorithms, and the automatic analysis of 5000 glossary terms across agencies (EIA, Census SICS and NAICS codes, EPA) towards the creation of a cross-agency ontology

User Interface Development. We have designed and implemented a completely new user interface with the capability of handling integrated querying and presentation of results.

2.2 EDC System Architecture

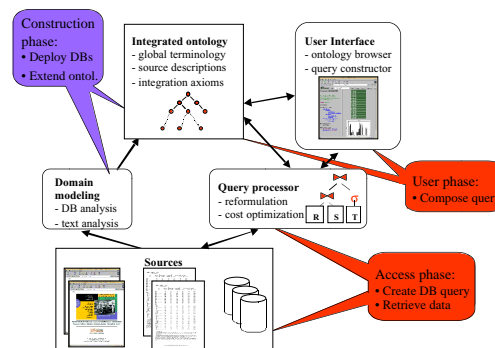


Figure 1. Architecture of EDC System.

In order to support homogeneous access to multiple databases, the EDC system involves three principal components: the database manager and access planner (Section 1.2), the overarching ontology in which terms are defined and standardized (Sections 1.3 to 1.5), and the interface (Section 1.6). The system architecture is shown in Figure 1.

2.3 The SIMS Planner and Domain Models

The retrieval of information dispersed among multiple sources requires familiarity with their contents and structure, query languages and location. A person (or system) with need for distributed information must ultimately break down a retrieval task into a collection of specific queries to databases and other sources of information (e.g., analysis programs). With a large number of sources, individuals typically do not possess the knowledge or time required to determine how to find and process the information they need. Even if they did, performing the necessary tasks would be tedious, time consuming, and prone to error.

Our approach to integrating statistical databases builds on research performed by the SIMS group at ISI (Arens et al., 1996). For several types of underlying data sources SIMS can translate a high-level request into a *query plan* (Ambite and Knoblock, 1997)—a series of operations including queries to sources of relevant data and manipulations of the data. Queries to SIMS are expressed internally in the Loom knowledge representation language (MacGregor, 1990). Each information source is described (*modeled*) in Loom, and an automatic planning system produces a query-plan based on these descriptions and its model of the application domain. Prior work on SIMS is being extended to deal with issues that are of particular significance in, statistical databases.

SIMS currently accepts queries in a subset of the SQL language against a high-level view of available data. The subset supported is limited in its treatment of aggregation operators. The problem is that distributing such operators over multiple databases is difficult and potentially inefficient. For example, finding the average of a distributed dataset is done fastest by retrieving only the average value and number of instances for each database and then calculating the global result—thereby minimizing transfer of data. However, if one of the DBMSs does not support averaging, all instances will have to be obtained

from it and the averaging done at the integration site. Obviously, it is better not to obtain so much data unnecessarily. Work on the aggregation problem is proceeding both at Columbia (Gravano et al., 1999; in prep) and at ISI (Ambite et al., 1997).

Since starting in 1999, the database portion of the EDC project has incorporated over 100 tables, from databases in various formats (Microsoft Access, HTML webpages, PDF, etc.), collected from the Energy Information Administration, the Census Bureau, the Bureau of Labor Statistics, and the California Energy Collection. Each table was ‘wrapped’ by the creation of specific access functions, attached to a data model, which in turn is mapped to a so-called domain model (represented in Loom) of the information contained. The SIMS access operators use the domain model’s terms to determine where pertinent information resides and to plan the access. A set of approx. 500 domain terms, represented in 10 subhierarchies, constitutes the domain models so far required for the EDC domain. These domain models describe such parameters of the data as *Area* (city, county, state, region, country, etc.), *Measuring Unit* (of oil, money, time, etc.), *Gasoline Type*, and so on. As described below, these models have been linked into the overarching SENSUS ontology.

2.4 The SENSUS Ontology

Practical experience has shown that integrating different termsets and data definitions is fraught with difficulty (Klavans and Wacholder, 1990). The U.S. Government has funded several meta-data initiatives with rather disappointing results. The focus has been on collecting structural information (formats, encodings, links), instead of content, resulting in large data collections (up to 500,000 terms) that are admirably neutral, but unsuitable as ‘terminology brokers’.

We are following a different approach, one that has been tested, on a relatively small scale, in various applications in the past two years. Rather than mapping between domains or collecting meta-data, we will create mappings between the domain and an existing *reference ontology*. This choice allows us in the future to make available to statistics agencies (and eventually to the general public) any other domains that have also been mapped into the reference ontology. Furthermore, by making

publicly available the reference ontology with our merging tools, we hope to encourage others to align (or even to merge) their termbanks, data dictionaries, etc., as well.

We are collecting, aligning, and merging the contents of several large termbanks, placing them under the high-level structure of an existing large and fairly general ontology called SENSUS, built at ISI. SENSUS (Knight and Luk, 1994) currently contains approx. 90,000 terms, linked together into a subsumption (*is-a*) network, with additional links for part-of, pertains-to, and so on. SENSUS is a rearrangement and extension of WordNet (Fellbaum, 1998) (built at Princeton University on general cognitive principles), retaxonomized under the Penman Upper Model (Bateman et al., 1989) (built at ISI to support natural language processing). For most of its content, SENSUS is identical to WordNet 1.5.

SENSUS can be accessed publicly using the Ontosaurus browser (Swartout et al., 1996), at http://mozart.isi.edu:8003/sensus/sensus_frame.html. SENSUS has been used to serve as the internal mapping structure (the Interlingua termbank) between lexicons of Japanese, Arabic, Spanish, and English, in several projects, including the GAZELLE machine translation engine (Knight et al., 1995), the SUMMARIST multilingual text summarizer (Hovy and Lin, 1999), and the MuST multilingual text retrieval and management system (Lin and Hovy, 1999). The GAZELLE and MuST lexicons contain over 120,000 root words (Japanese), 60,000 (Arabic), 40,000 (Spanish), and 90,000 (English), and 90,000 (Bahasa Indonesia), of which various amounts have been linked to SENSUS. SENSUS terms serve as connection points between equivalent language-based words.

A version of SENSUS has been deployed for the EDC project.

2.5 Semi-Automated Term-to-Ontology Alignment

In linking agency-specific domain models (as required by SIMS) to SENSUS (and hence to one another), the central problem is (semi-) automated term alignment and merging. Determining where a given term belongs in a 90,000-node ontology is not trivial! At first glance, it might seem impossible to align two ontologies (or taxonomized termsets) automatically. Almost all ontologies, after all,

depend to a large degree on non-machine-interpretable information such as concept/term names and English term definitions. However, recent research has uncovered a variety of heuristics that help with the identification and alignment process. We are using a 5-step procedure that is partially automated:

- heuristics that make initial cross-ontology alignment suggestions
- a function for integrating their suggestions
- a set of alignment validation criteria and heuristics
- a repeated integration cycle
- an evaluation metric

The full power of these techniques is still being explored, either linking words from foreign lexicons (Knight and Luk, 1994; Okumura and Hovy, 1994) or concepts from other ontologies (Ageno et al., 1994; Rigau and Agirre, 1995; Hovy, 1998). In the EDC project we have implemented two old matching heuristics (NAME and DEFINITION MATCH) and developed a new one (DISPERSAL MATCH) to link both approx. 100 domain model concepts used for SIMS and approx. 6000 glossary items acquired from the EIA (see below) into the ontology. As explained in (Hovy et al., 2000), the accuracy of linkage (for the domain model concepts) is correlated to the degree of dispersal of the target concepts within SENSUS. Figure 2 illustrates the accuracy for each of ten subgroups of domain concepts, measured against human alignment of the same concepts. The almost-perfect accuracy of 8 of the 10 subgroups is cause for optimism.

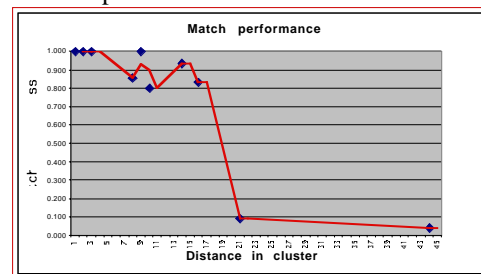


Figure 2. Accuracy of alignment of 98 domain model concepts (organized in 10 groups) against SENSUS (90,000 concepts), using automated alignment heuristics.

Current research is focusing on efficient implementation of the matching procedure (the initial process took over 28 hours for fully aligning approx. 100 concepts into SENSUS).

2.6 Extracting Glossary Entries with LKB

Several problems with terminology require that special attention be devoted to terms in a cross-agency statistical data provision system. In particular, and especially confusing to non-specialist users, is the proliferation of terms in the domain, and the fact that agencies define terms differently, even when (ostensibly) relating to the same object or fact. What is called *wages* in one database may be what another calls *salary* (even though it may have a *wages* too, as well as *income*). Reading the glossaries' definitions may or may not be a help; lengthy term definitions often contain important information that is buried.

The Lexical Knowledge Base (LKB) system is being developed at Columbia University to extract information from glossaries and output it in a form appropriate for inclusion in SENSUS, and eventually for automated processing; see (Klavans and Whitman, 2000). This work is based on prior experience on dictionary analysis to create a knowledge base (Klavans, 1988), automatic phrase variation rules for mapping related terms (Klavans et al., 1997), the use of lexical resources to determine the range of ambiguity, verification in corpora to confirm ambiguity measures, and the use of linked phrases to eliminate potential ambiguity of single word terms. LKB combines statistical and linguistic methods in finite-state patterns to identify topics with high accuracy and decompose their definitions into frames with several slots, including head term, genus term (*isa*) cross-refs, items included and excluded, and so on. LKB is designed to provide complete coverage and be useful for any subject area; it has produced over 6,000 concepts in the current EDC domain.

At present, the output frames are filled by pure extracts (text) from the glossary. Current work is focusing on converting and systematizing some of this content into expressions in a formal language, suitable for reasoning.

2.7 EDC Interface

It is currently extremely difficult for users to make productive use of the statistical data available on the web. Because of the sheer wealth of information, current systems typically offer two fundamentally different user interfaces to access it. One method attempts to trade off

generality for ease of use by relying on a collection of ready-made presentations, consisting of tables and charts that have been designed in advance to answer typical questions. However, there may be hundreds of these presentations, making it difficult for users to find the one that provides the closest answer. At best, these systems provide a keyword searching mechanism to help users discover relevant presentations; in many cases, none of them may address the user's specific query.

The other method for finding information achieves generality by allowing users to construct their own queries. However, these user interfaces are for experts only, requiring intimate knowledge of the domain and structure of the database, the meaning of the attributes, the query language, and the ways in which resulting information is presented.

To address the user interface issues, we are developing a unified web-based user interface for querying and presenting statistical information. This interface allows people to express queries using an adaptive forms-based method. This method allows users to construct complete queries by choosing from a dynamically changing set of menu options, composed dynamically with reference to the domain models in SENSUS. The design is obviously extensible: as new databases are added to the system, their domain models are linked into SENSUS, and their parameters are immediately available to the user for querying. The taxonomization in SENSUS ensures appropriate grouping for menu display by the interface.

The interface, which is being constructed by Columbia University, communicates with SIMS and SENSUS via an API specially developed to support appropriate data transfer.

3 Conclusion: Future Work

In a recently submitted proposal, we requested funding to extend the EDC project in three significant directions:

1. Multilingual access of information. We propose to extend the basic system from English-only to include Spanish, Chinese (Mandarin, and Taiwanese if possible), and a third language, to accommodate the large Spanish and other linguistic groups present in both urban and rural areas throughout the United States. We will build upon the considerable resources and

expertise in multilingual language processing of the Natural Language Group at ISI.

2. Efficient and effective techniques high speed query evaluation. In order to further decrease the response time of complex queries, we will study new query processing methods for main-memory databases at Columbia University. With gigabytes of main memory now cheaply available, it is often feasible for analysts to load their datasets of interest into a local main-memory database in order to pose a variety of complex analytical queries.

3. Formal evaluation of system performance. Both for the expert and for the novice, obtaining information from government data sources can be daunting. In order to deliver better service to users, we propose to experimentally determine the effectiveness of various interfaces, interaction paradigms, and system behaviors, in order to guide further research and system development. We will employ the EDS testing service at Columbia University.

4 References

- Ageno, A., I. Castellon, F. Ribas, G. Rigau, H. Rodriguez, A. Samiotou. 1994. TGE: Tlink Generation Environment. *Proceedings of the 15th COLING Conference*. Kyoto, Japan.
- Ambite, J.-L. and C.A. Knoblock. 1997. Planning by Rewriting: Efficiently Generating High-Quality Plans. *Proceedings of the 14th National Conference on Artificial Intelligence*. Providence, RI.
- Ambite, J.-L., Y. Arens, L. Gravano, V. Hatzivassiloglou, E.H. Hovy, J. Klavans, A. Philpot, J. Sandhaus, A. Singla, B. Whitman. 2000. The EDC Project. *Report of the First NSF Conference on Digital Government*. Los Angeles, May 2000. In prep.
- Arens, Y., C.A. Knoblock and C.-N. Hsu. 1996. Query Processing in the SIMS Information Mediator. In A. Tate (ed), *Advanced Planning Technology*. Menlo Park: AAAI Press.
- Bateman, J.A., Kasper, R.T., Moore, J.D., and Whitney, R.A. 1989. A General Organization of Knowledge for Natural Language Processing: The Penman Upper Model. Unpublished research report, USC/Information Sciences Institute, Marina del Rey, CA.
- Fellbaum, C. 1998. (ed.) *WordNet: An On-Line Lexical Database and Some of its Applications*. Cambridge: MIT Press.
- Gravano L., H. Garcia-Molina, A. Tomasic. 1999. GLOSS: Text-Source Discovery over the Internet. *ACM Transactions on Database Systems*.
- Gravano, L. in prep.
- Hovy, E.H. 1998. Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses. *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*. Granada, Spain.
- Hovy, E.H. and C.-Y. Lin. 1999. Automated Text Summarization in SUMMARIST. In M. Maybury and I. Mani (eds), *Advances in Automatic Text Summarization*. Cambridge: MIT Press.
- Hovy, E.H., A. Philpot, J.-L. Ambite, and U. Ramachandran. 2000. Automating the Placement of Database Concepts into a Large Ontology. In prep.
- Klavans, J.L. 1988.
- Klavans, J.L. et al., 1997.
- Klavans, J.L. and Wacholder, 1990.
- Klavans, J.L. and B. Whitman. in prep.
- Knight, K. and S.K. Luk. 1994. Building a Large-Scale Knowledge Base for Machine Translation. *Proceedings of the AAAI Conference*.
- Knight, K., I. Chander, M. Haines, V. Hatzivassiloglou, E.H. Hovy, M. Iida, S.K. Luk, R.A. Whitney, and K. Yamada. 1995. Filling Knowledge Gaps in a Broad-Coverage MT System. *Proceedings of the 14th IJCAI Conference*. Montreal, Canada.
- Lin, C.-Y. and E.H. Hovy. 1999. The MuST System. *Proceedings of the 22nd SIGIR Conference*, Berkeley, CA.
- MacGregor, R. 1990. The Evolving Technology of Classification-Based Knowledge Representation Systems. In John Sowa (ed.), *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. Morgan Kaufmann.
- Okumura, A. and E.H. Hovy. 1994. Ontology Concept Association using a Bilingual Dictionary. *Proceedings of the 1st AMTA Conference*. Columbia, MD.
- Rigau, G. and E. Agirre. 1995. Disambiguating Bilingual Nominal Entries against WordNet. *Proceedings of the 7th ESSLI Symposium*. Barcelona, Spain.
- Swartout, W.R., P. Patil, K. Knight, and T. Russ. 1996. Toward Distributed Use of Large-Scale Ontologies. *Proceedings of the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop*. Banff, Canada.