

**2000: Project Report to NSF**

**Energy Data Collection Project**

*Year 1*

**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 Amit Singla† and Brian Whitman†**

\* Digital Government Research Center  
Information Sciences Institute  
University of Southern California  
4676 Admiralty Way  
Marina del Rey, CA 90292-6695

Digital Government Research Center  
Department of Computer Science  
Columbia University  
535 West 114th Street, MC 1103  
New York, NY 10027

**Contacts:** Eduard Hovy, ISI  
hovy@isi.edu

Judith Klavans, Columbia  
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. At the Digital Government Research Center 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.

• **Introduction: The Digital Government Research Center**

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 dg.o2000, the first workshop of the National Science Foundation's Digital Government program, in Los Angeles in May 2000 (see <http://www.drcc.org>).

- **The EDC Project**
- **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 short 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, performed research on computational properties of aggregation, and investigated the extraction of information from footnotes embedded in text. See Section 3.

**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 domain models and extending and developing new automated concept-to-ontology alignment algorithms. Term extraction from glossaries involves the automatic analysis of 5000 terms across agencies (EIA, Census SICS and NAICS codes, EPA) and the automatic handling of acronyms towards the creation of a cross-agency ontology. See Section 4.

**User Interface Development.** We have designed and implemented a completely new user interface with the capability of handling integrated querying and presentation of results. See Section 5.

- **EDC System Architecture**

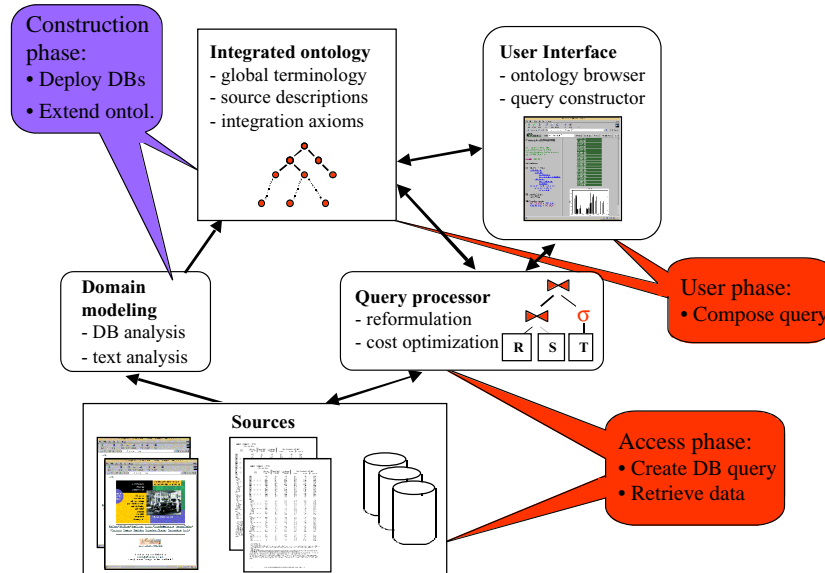


Figure 1. Architecture of EDC System.

In order to support homogeneous access to multiple databases, the EDC system involves three principal components: the database manager and access planner, the overarching ontology in which terms are defined and standardized, and the interface. The ontology construction phase at top right includes the work on semi-automated term alignment, term extraction from glossaries, and acronym handling. The system architecture is shown in Figure 1.

- **Information Capture and Integration**

In this section we discuss the issues of database wrapping, domain modeling, the SIMS data access planner, query aggregation, and the extraction of information from footnotes.

- **The SIMS Planner and Domain Models**

Our approach to integrating statistical databases builds on research performed by the SIMS group at ISI (Arens et al., 1996). SIMS assumes that a set of information sources such as databases, web servers, etc., supply data about a particular application domain. The system designer specifies a global model of the application domain and describes the contents of the sources in terms of this global model. A SIMS mediator integrates and provides a single point of access for all the information in such a domain. The user interacts directly with the SIMS mediator expressing queries against the domain model, without needing to know about the schemas or locations of the multiple sources.

SIMS translates a high-level request into a *query plan* (Ambite and Knoblock, 2000)—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. The subset supported is limited in its treatment of aggregation operators (such as sum, average, etc.). 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. We discuss our approach for queries involving aggregation operators below.

- **Incorporating and Modeling New Databases**

Since starting in 1999, we<sup>1</sup> have incorporated over 100 tables, from sources in various formats, (including Oracle and Microsoft Access databases, HTML web forms and pages, and PDF files), collected from the Energy Information Administration, the Census Bureau, the Bureau of Labor Statistics, and the California Energy Commission.

A large amount of the information is in the form of semi-structured web pages. These web sources were ‘wrapped’ automatically using technology from the Ariadne system (Muslea et al., 1998). Ariadne allows a developer to mark up example web pages using a demonstration-based GUI. Then the system inductively learns a landmark grammar that is used to extract the marked-up fields from similar pages and generates all the necessary wrapper code. The resulting wrapper acts as a simple relational database that accepts parametrically-defined SQL and dynamically retrieves data from the associated web pages and forms.

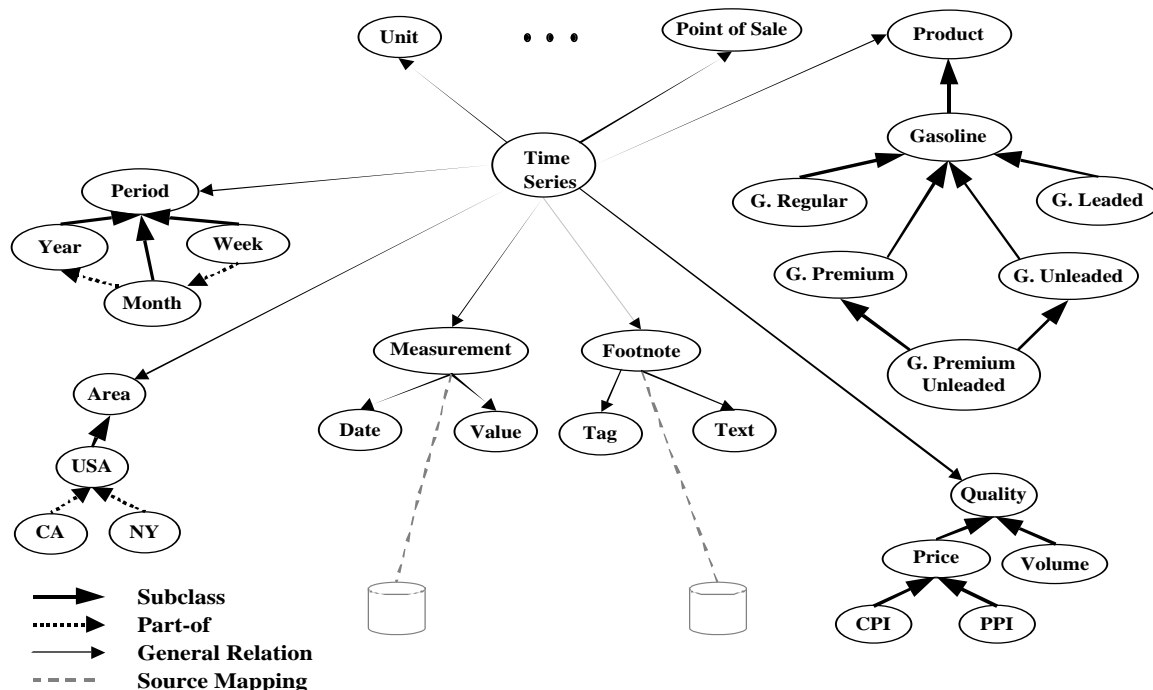


Figure 2. Fragment of the Domain Model.

In SIMS, each of these data sources, whether natively relational or wrapped by Ariadne, is modeled by associating it to an appropriate domain-level concept description. A set of approximately 500 domain terms, organized in 10 subhierarchies, constitutes the domain model so far required for the EDC domain. A fragment of the EDC domain model is shown in Figure 2. This model describes time series data about different gasoline products. A time series is defined by a set of dimensions such as *product type* (e.g., unleaded gasoline, premium gasoline), *property measured* (e.g., price, volume), *area* of the measure (e.g.,

<sup>1</sup> This work was performed by Andrew Philpot and José Luis Ambite, of USC/ISI, with the help of domain experts at EIA and BLS.

USA, California), *unit* of measure, etc. Each of the time series in the sources is described by using specific values for each of the hierarchical dimensions. For example, a particular source may be described as providing the monthly prices (based on the consumer price index) of premium unleaded gasoline for the state of California. The dimensions can be seen as metadata that describes the series. The actual data is modeled as a set of measurements (i.e., date and value pairs). The domain model also describes whether a source has footnotes for some of the data. The answer to a query will also return the footnote data associated with the corresponding tuples if so requested.

These models have been linked into the overarching SENSUS ontology. Each of the retrievable time series, along with each of the ten dimensional values, has been added to SENSUS as an ontological concept in its own right; the relationships between series and dimensional values have been reified as SENSUS relations as well (e.g., has-product-type, area-of, etc.). In Section 4.1 we discuss how this linking was performed semi-automatically. Using tools that facilitate the construction of wrappers and the semi-automatic description of sources is critical to scale mediator systems to the very large number of information sources that are available from government agencies in a cost-effective fashion.

- **Query Aggregation**

We<sup>2</sup> have been investigating how to integrate data sets/sources with information aggregated at different ‘granularities’ and with different ‘coverage’. For example, a data source might have gasoline-price information for the whole United States reported by month for the last ten years; another source might have the same type of information for the whole United States reported by year up to 1990; finally, yet another source might have yearly gasoline-price information discriminated by state. The goal of our work is to conceptually present users with a reasonably uniform view of the available data without necessarily exposing all this heterogeneity in aggregation granularity and coverage.

The main challenge of our integration is dealing with data sets exhibiting *varying* granularity and coverage. In effect, data sets might have information at different granularities of time (e.g., month vs. year), of geography (e.g., cities vs. states vs. countries), of product (e.g., unleaded regular gasoline vs. ‘general’ gasoline), and so on. For example, data sets might have information with different coverage in terms of time (e.g., January 1978 through December 1986 vs. January 1978 through January 1989), geography (e.g., San Diego, CA vs. Boston, MA), product (e.g., leaded premium gasoline vs. leaded regular gasoline), and so on. Our approach is to present users with a simple, unified view of the data. This view should be sufficiently fine grained so that users can exploit most of the information that is available, but it should also be sufficiently coarsely grained so that we hide most of the granularity and coverage differences of the data sets from users. After defining such a view, users pose queries against it. Most of the time we will be able to answer correctly user queries with the available data sets. Unfortunately, some other times we might have to reformulate a user query if the data that is needed to answer it is not available.

In case a user poses a query that cannot be answered with the available data, our approach is to relax and reformulate the user query. We find the *closest query* for which we have all required data, and provide exact answers for this closest query. Our key observation for this reformulation is that data attributes (e.g., time, geography, product) often follow natural granularity hierarchies (e.g., day->month->year for time, city->state->country for geography, leaded gasoline; unleaded gasoline->gasoline for product). Combining each of these granularity hierarchies results in a granularity *lattice* where a node might correspond to leaded gasoline data by month and state, and another node might correspond to leaded gasoline data by year and country, etc. We have developed algorithms to identify, for each of these nodes, the queries that we can answer exactly at the node’s level of granularity. For example, given a set of data sources, we might conclude that we can answer any query on leaded gasoline by month and state as long as the query is about the 1990–1999 time period, and about the states of California and New York. At run time, we can then decide whether we can answer a user query exactly as is, or whether we need to reformulate it and find the ‘closest’ query in the lattice for which we can provide exact answers, using some distance function over the granularity lattice. To illustrate our initial results and algorithms, we have developed a demo system that

---

<sup>2</sup> This work is led by Prof. Luis Gravano, an assistant professor at Columbia University, in collaboration with Vasilis Vassalos, an assistant professor at New York University, and Anurag Singla, a first-year Ph.D. student.

uses four BLS data sets, all on average price of unleaded regular gasoline. The demo is accessible at <http://db-pc01.cs.columbia.edu/digigov/Main.html>. See also (Gravano, 2000).

- **Automatic Footnote Extraction**

Footnotes are an important piece of metadata that often accompanies statistical tables. Footnotes may qualify the data of the whole table, a particular column, or specific cells in a table. Defining general procedures for the extraction of footnotes and determining the scope of applicability of the footnotes is a very challenging problem when the statistical tables come from text or HTML documents as is the case in much of the available government data.

We<sup>3</sup> have performed foundation research on the topic of automatically extracting footnotes and links between footnotes and text from web pages and tables. We have built finite-state analyzers that track the extent of each footnote and associate footnote symbols with footnote text. This will enable the fully automatic recognition of footnotes in tables in the future, and become the basis for textual analysis of footnotes to detect differences between concepts or sources.

- **Ontology Construction**

In this section we discuss the issues of cross-agency terminology standardization using the SENSUS ontology. This includes the semi-automated linking of new terms into SENSUS, the extraction of terms from agency glossaries, and the handling of acronyms.

- **The SENSUS Ontology**

Practical experience has shown that integrating different termsets and data definitions is fraught with difficulty. 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'.

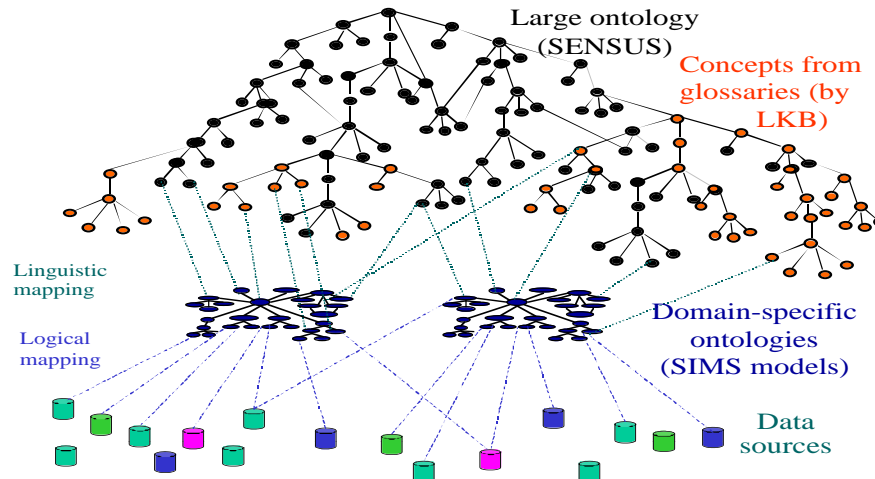


Figure 3. Ontology and Domain Models.

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

---

<sup>3</sup> This work was performed by Vasilis Hatzivassiloglou, from Columbia University, with the assistance of a graduate student, Jay Sandhaus.

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.

The ontology for the EDC project has the structure shown in Figure 3. The bulk of the ontology comes from SENSUS augmented with the SIMS domain model and concepts automatically extracted from online glossaries. To deploy the version of SENSUS used for the EDC project, we<sup>4</sup>: (1) created a new copy of the definition files of the whole ontology, (2) created a new copy of the Ontosaurus browser and installed it on <http://ariadne.isi.edu:8005/sensus-edc/>, (3) defined a domain model consisting of approximately 500 nodes to represent the concepts present in the EDC gasoline domain (see Section 3.2), (4) linked these domain concepts into SENSUS using semi-automated alignment tools (see Section 4.2), and (5) defined a new type of ontology link called *generally-associated-with* to hold between concepts in the ontology and domain model concepts, allowing the user while browsing to rapidly proceed from high-level concepts to the concepts associated with real data in the databases.

SENSUS can be accessed publicly using the Ontosaurus browser (Swartout et al., 1996), at [http://mozart.isi.edu:8003/sensus/sensus\\_frame.html](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. To deploy version of SENSUS for the EDC project, we<sup>5</sup> did the following:

- we created a new copy of the definition files of the whole ontology,
- we created a new copy of the Ontosaurus browser, and installed it on <http://ariadne.isi.edu:8005/sensus-edc/>,
- we defined a domain model consisting of approx. 500 nodes to represent the concepts present in the EDC gasoline domain (see Section 3.2),
- we linked these domain concepts into SENSUS using the alignment tools (Section 4.2),
- we defined a new type of ontology link called *generally-associated-with* to hold between concepts in the ontology and domain model concepts, allowing the user while browsing to rapidly proceed from high-level concepts to the concepts associated with real data in the databases.

#### • **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.

---

<sup>4</sup> This work was performed by José Luis Ambite, Eduard Hovy, and Andrew Philpot, of USC/ISI.

<sup>5</sup> This work was performed by Jose-Luis Ambite, Eduard Hovy, and Andrew Philpot, of USC/ISI.

However, recent research has uncovered a variety of heuristics that help with the identification and alignment process. We<sup>6</sup> 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 re-implemented two existing matching heuristics (NAME and DEFINITION MATCH) and developed a new one (DISPERSAL MATCH). NAME MATCH performs an exhaustive (sub)string match of the concept name to be linked against every concept name in SENSUS, with special rewards for beginning and ending overlaps of substrings. Since this match is very slow, we have implemented an algorithm used to match gene sequences to obtain a two order of magnitude speedup. DEFINITION MATCH compares the overlap of words in the definition of the concept to be linked against the definitions of all SENSUS concepts, after appropriate demorphing and closed-class word removal. We have implemented a standard IR-based vector space matching algorithm for an efficient implementation.

DISPERSAL MATCH is a new invention, developed for this project (Hovy et al., 2000). This heuristic is based on the expectation that a set of concepts to be linked, if they are semantically related, will cluster together inside SENSUS and not be widely dispersed, given that SENSUS concepts are also organized by semantic closeness. It turns out that this heuristic performs rather well. We have applied it to link the approx. 100 domain model concepts used for SIMS into SENSUS, organized into 10 subgroups, and are at the time of writing linking the approx. 6000 glossary items acquired from the EIA (see Section 4.3) as well. As is to be expected, the accuracy of linkage is correlated to the degree of dispersal of the target concepts within SENSUS. Figure 4 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 provides cause for optimism.

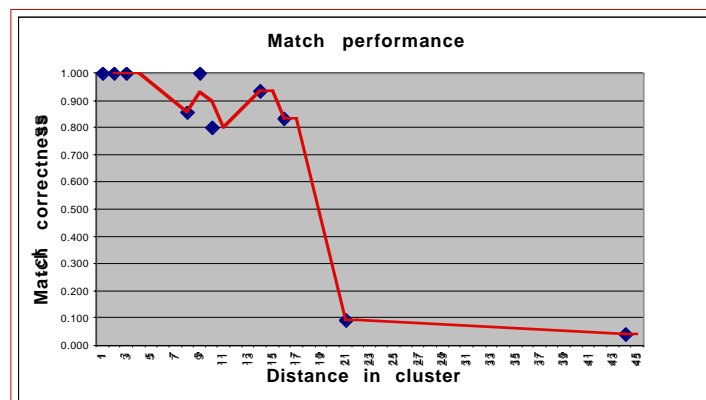


Figure 4. Accuracy of alignment of 98 domain model concepts (organized in 10 groups) against SENSUS (90,000 concepts), using automated alignment heuristics.

- **Definition Analysis: 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

<sup>6</sup> This work was performed by Eduard Hovy of USC/ISI, with a graduate student, Usha Ramachandran. Refinements and recoding of the algorithms are being performed by Andrew Philpot of USC/ISI.

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 goal of the first year's research<sup>7</sup> on definition analysis was to identify a set of resources across agencies relevant to the domain of energy data, to develop tools to automatically parse these definition sets, regardless of their internal complexity, and to map them into a lexical knowledge base of uniform structure. From this representation, a mapping into ISI's SENSUS can be performed, at first semi-automatically with increasing levels of automation as each component is improved. The primary challenge of year one was to identify the relevant sources, to build the definition parser, and to attempt to link related phrases via their semantic components.

The Columbia Automatic Lexical Knowledge Base (ALKB) system takes a definition source (web page or document) and creates a Lexical Knowledge Base automatically. An LKB is a structured form of a set of definitions, and can be used for ontology generation and definition analysis. 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.

The system uses a combination of rule-based techniques enhanced with statistical methods. Part of speech tagging is used to analyze the definition, and then LinkIT, a noun-phrase chunker from Columbia University, is applied to determine linked noun phrases. At the same time, a bigram analysis is performed to determine potential collocational attributes. From these two methods, various semantic attributes are tagged comprising two types: (1) predefined semantic attributes which are determined after an analysis of definition literature and a definition set. These include such attributes as "contains," "used for," "excludes," "includes," and so on. These are arranged into three categories: properties, excludes/includes, and quantifiers. (2) automatically determined potential attributes: determined after running the bigram probability model across the entire document to find other attributes that might be useful in classifying the document. A group of these are identified and, if they occur in the definition currently being analyzed, are shown under the analysis with the phrases surrounding it. This way, a user can note which attributes might need to be added to the already predefined set. Output from predefined semantic attribute analysis is used to build a list-like representation for input into SENSUS. Output from probabilistic analysis is meant for the user to post-process.

In the first year, ALKB was used to analyze a total of nearly 8000 definitions from the following sources: (1) nineteen (19) long description definitions from the official EIA Gasoline Glossary; (2) 2588 definitions from EIA's larger glossary set (unedited); (3) an additional 125 definitions from EIA's larger set; (4) 4526 definitions from EPA's Glossary of Selected Terms and Abbreviations; and (5) about 35 relevant SIC and NAICS code metadata definitions and explanations from the Census Bureau. In addition, ALKB was run over the output of an automatic extraction of medical definitions (91 terms and definitions) from lay articles in the Digital Library II project at Columbia. The ALKB web page shows the system: <http://www.cs.columbia.edu/nlp/flkb/>.

In the next year, our goal is to work with the ontology team at ISI to ensure that output of ALKB can be more automatically used as input to SENSUS, since there are still mapping issues due to the complexity of the data. The mapping issues involve interglossary merging issues as well as external mapping into SENSUS. In addition, we will further perfect the parser to obtain more refined results. We will seek additional glossary in the energy domain, perhaps using definitions extracted from text.

- **Acronym Analysis with Acrocat**

The ALKB system uses the Acrocat acronym cataloguing system to try to determine the meaning of acronyms used in the document. A list of possibilities for each acronym in the current definition are listed

---

<sup>7</sup> This work was led by Judith Klavans at Columbia University, with a graduate student, Brian Whitman.

along with a confidence marker. Acrocat was developed<sup>8</sup> as a sub-routine of ALKB since agency-specific abbreviations and acronyms are frequent in definitions and thus often make these definitions uninterpretable outside a given agency or domain. Code for initial acronym resolution was built in year one. In year two, we will link Acrocat with existing acronym and abbreviation glossaries in order to add guesses from these external resources. One of the most challenging problems in dealing with acronyms is to resolve ambiguity, since the determination of the expansion of an acronym is often domain-dependent (e.g. NFS could be Not For Sale in the art or auction worlds, but refers to Network File System in computer science). Currently, Acrocat has a demo page, which permits a user to enter source data directly, at <http://www.cs.columbia.edu/nlp/acrocat/>.

- **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<sup>9</sup> are developing a unified web-based user interface for querying and presenting statistical information. Our focus in the first year of the project has been on the development of a robust, portable, and efficient user interface that facilitates user access to data from multiple sources/agencies. The interface addresses the following main tasks: support for adaptive, context-sensitive queries via a system of guided menus; display of tables created by the integration back-end from one or multiple individual databases, along with footnotes and links to original data sources; and browsing of the ontology that supports the entire integration model, with the capability to display concept attributes, relationships, and definitions in graphics and text. 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 was implemented as a thin client in Java/Swing, in a way that provides a professional look-and-feel and still runs within any web browser with a minimal download. This will allow even casual users to obtain the downloadable application interface from our server. All text analysis, ontology support, and database integration management is performed at DGRC servers in a manner transparent to the users. The interface client has been tested on Unix, Windows, and Mac operating systems. The interface communicates with SIMS and SENSUS via an API specially developed to support appropriate data transfer.

We plan to extend our user interface with additional components that visualize data with graphics and add personalization capabilities that track a given user's queries. We will also add support for complex queries on aggregated data or with queries on partially unspecified information. We plan an initial user study of the effectiveness of our system, with support from the Bureau of the Census and Columbia's Electronic Data Center. We will also continue, and gradually focus more on, our research-oriented work with footnotes, developing ways to analyze the scope of footnote attachments and means for locating the principal concept that a footnote refers to.

---

<sup>8</sup> This work was led by Judith Klavans of Columbia University and a graduate student, Brian Whitman.

<sup>9</sup> This work was performed by Vasilis Hatzivassiloglou of Columbia University, with a graduate student, Jay Sandhaus.

- **Conclusion**
- **Future Work**

We have presented an approach to integrate and facilitate access to heterogeneous statistical information sources that different government agencies provide. We use a large ontology to facilitate the integration and user access. We have developed query processing techniques that take into account particular characteristics of statistical government information such as the presence of footnotes and data aggregated at different granularities. Our pilot system integrates data from web sources and databases from several federal agencies such as the Bureau of Labor Statistics, the Energy Information Administration, and the Census Bureau.

We intend to expand our system in several directions. First, we plan to tailor the interface for different levels of user expertise. For naïve users we plan to leverage SENSUS to recognize some simple queries formulated in limited natural language, in addition to the capability of browsing the ontology that is currently available. For more expert users we will allow the specification of increasingly complex queries expressed in terms of the domain model.

Second, we will focus in efficient query processing. We will investigate improved algorithms for query reformulation in the presence of aggregation operators and complex analytical queries. In order to further decrease the response time of complex queries, we will study new query processing methods for main-memory databases. 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.

Finally, we will explore further techniques to facilitate the addition of new sources to the system, the construction of domain models, and their incorporation into the overarching ontology, with a minimum of manual intervention. In order to add a new source the designer needs to understand its semantics and describe it using the appropriate terms from the domain model. We plan to analyze both the data in the sources to obtain constraints that describe it (for example, that the data in a source is between 1960 and 1990) and the available metadata such as related text, glossaries, footnotes, table names, columns, and data values. We expect that using a large ontology based on natural language principles like SENSUS will allow us to semi-automatically assign semantics to the sources using similarity and clustering techniques on the available metadata.

- **Extending the System beyond Current Plans**

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.

## • 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., Y. Arens, L. Gravano, V. Hatzivassiloglou, E.H. Hovy, J.L. Klavans, A. Philpot, U. Ramachandran, J. Sandhaus, A. Singla, B. Whitman. 2000. Building Ontologies and Integrating Data from Multiple Agencies: A Case Study Using Gasoline. Paper 1941, *2000 Joint Statistical Meetings*. Indianapolis, IN.
- Ambite J.L. and C.A. Knoblock. 2000. Flexible and Scalable Cost-Based Query Planning in Mediators: A Transformational Approach. *Artificial Intelligence Journal*, 118 (1-2).
- 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. 2000. in prep.
- Harinarayan, V., A. Rajaraman, and J. D. Ullman, 1996. Implementing Data Cubes Efficiently, *Proceedings of the 1996 ACM SIGMOD Conference*.
- 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. COMPLEX: A Computational Lexicon for Natural Language Processing. *Proceedings of Twelfth International Conference on Computational Linguistics (COLING)*. Budapest, Hungary.
- Klavans, J. L., C. Jacquemin and E. Tzoukermann. 1997. "A Natural language approach to multi-word term conflation". *Proceedings of the DELOS conference* from the European Research Consortium on Information Management (ERCIM). Zurich, Switzerland.
- Klavans, J. L. and Muresan S. 2000 (in press). "DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text". *Proceedings of 2000 American Medical Informatics Association (AMIA) Annual Symposium*, Los Angeles, California.
- 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 22<sup>nd</sup> 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.