Simplifying Data Access: The Energy Data Collection Project

Using technology developed at the Digital Government Research Center, a team of researchers seeks to make government statistical data more accessible through the Internet.

José Luis Ambite
Yigal Arens
Eduard Hovy
Andrew Philpot
University of Southern California/ Information Sciences Institute

Luis Gravano
Vasileios Hatzivassiloglou
Judith Klavans
Columbia University

In a democratic society, the government has a mandate to make most information it collects available to the public. Responding to demands for data by statisticians, policymakers, researchers, businesses, investors, educators, and others, federal and state agencies are providing access to a vast amount of statistical data in electronic form. Making this information accessible and useful has posed two major challenges to the research and analysis communities. The first is integrating large, dispersed collections of data compiled by different people at different times and for different purposes. The second is overcoming the limitations of the Web's browser paradigm to disseminate complex information derived from multiple sites.

The Digital Government Research Center (DGRC) unites researchers and developers from the University of Southern California's Information Sciences Institute (ISI) and Columbia University's Department of Computer Science and its Center for Research on Information Access to address these problems (http://www.dgrc.org). In collaboration with government experts, we conduct research in advanced information systems; develop standards, interfaces, and a shared infrastructure; and build and manage pilot systems.

THE EDC PROJECT

The DGRC's Energy Data Collection (EDC) project began in 1999 as a pilot of the National Science Foundation's Digital Government Program, which funds research to meet the needs of federal information service communities. Working with representatives of federal and state statistics agencies and other organizations, the EDC is building a system for disseminating energy data from the Bureau of Labor Statistics (BLS), the Census Bureau, the Department of Energy's Energy Information Administration (EIA), and the California Energy Commission (CEC).

An example of the information the program provides is the extensive monthly energy statistics posted on the EIA's Web site (http://www.eia.doe.gov). Although it receives hundreds of thousands of hits each month, the site currently supports limited access to this very rich information source. Most data is available only for recent years and only as downloads of standard HTML pages or PDF documents. The many definitions and footnotes that explain the complexity of the data are not available, and the site's query definition facility is too difficult for anyone but experts to use.

In our first year of research and development, we demonstrated initial results in three areas:

- *Information integration*. We have developed effective methods to identify and describe the contents of databases, making it possible to accurately and efficiently locate useful information even when precise answers are unavailable. We have wrapped more than a hundred Web sources for the first stage of integration, performed research on computational properties of aggregation, and investigated how to extract information from footnotes embedded in text.
- *Ontology construction*. We have extended the USC/ISI 90,000-node Sensus terminology taxonomy to incorporate new energy-related domain models, and we have developed automated, concept-to-ontology alignment algorithms. We have created a cross-agency ontology that automati-

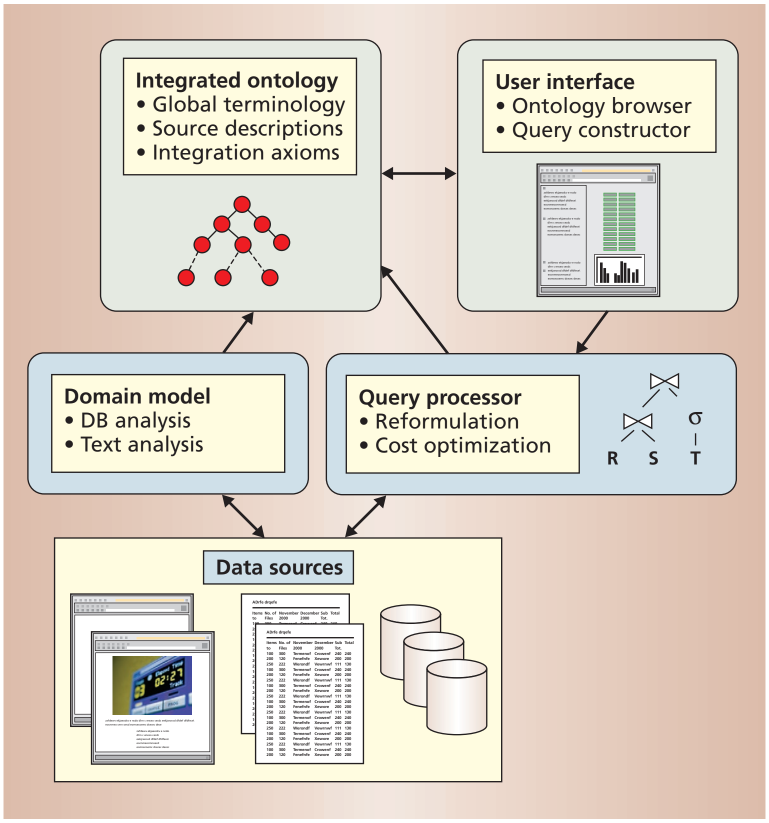

Figure 1. The EDC system architecture includes an integrated ontology, a user interface, a query processor, a domain model, and the data sources.

cally extracts 7,000 terms from multiple glossaries and analyzes acronyms.

- *User interface development.* We have designed and implemented a completely new user interface capable of integrating queries and presenting results.

As Figure 1 shows, the EDC system's architecture includes the following major components: an integrated ontology, a user interface, a query processor, a domain model, and the data sources. The ontology includes a high-level general concept taxonomy and links to the domain model, which models the contents of new databases and extends the ontology as required; this work involves semiautomated term alignment, glossary entry extraction, and acronym analysis.

The interface facilitates construction of user queries, which may involve ontology browsing and other interaction methods. The user interface dispatches a high-level query to the query processor, which returns results to the interface for display (if the available sources cannot answer a query exactly, the interface may engage the user in a query dialogue). To answer user queries, the query planner consults the ontology and domain model to transform the high-level user query into an optimized database query plan that accesses the relevant sources, retrieves the data, and composes the requested information.

INFORMATION INTEGRATION

Retrieving relevant data distributed among many sources at different agencies requires familiarity with the contents, structure, query languages, and location of various databases and analysis programs. Most of us do not possess the knowledge, time, or patience required to break down a retrieval task into a collection of specific queries from multiple heterogeneous information sources. One needs a single access mechanism to allow users to express queries without having to know anything about the individual sources.

SIMS query plans

Our approach to integrating statistical databases builds on work performed by ISI's SIMS (Single Interface to Multiple Sources),[1] a group of research projects investigating aspects of retrieving and integrating data distributed among multiple heterogeneous information sources. SIMS assumes that the system designer specifies a global model of the application domain and describes the contents of each source—databases, Web servers, and so on—in terms of this global model. A SIMS mediator provides a single point of access for all this information.

SIMS translates the user's high-level request, expressed in a subset of SQL, into a *query plan*.[2] Such a plan is a series of operations that includes queries to relevant data sources and data manipulations. Queries are expressed internally in the Loom knowledge representation language.[3]

There is a limit to the SQL subset's treatment of aggregation operators such as sum and average. Distributing such operators over multiple databases is both difficult and potentially inefficient. For example, the fastest way to find the average of a distributed data set is to retrieve only the average value and number of instances for each database and then calculate the global result, thereby minimizing data transfer. However, if one of the DBMSs does not support averaging, all instances will have to come from that database, and the program will have to average at the integration site. Obviously, it is better to avoid obtaining too much information unnecessarily.

Domain models

We have thus far incorporated more than 100 tables from the BLS, Census Bureau, EIA, and CEC in various formats, including Oracle and Microsoft Access databases, HTML Web forms and pages, and PDF files. Most of the information is in the form of semistructured Web pages, which we wrapped automatically using technology from ISI's Ariadne system.[4] Ariadne allows a developer to mark up example Web pages using a demonstration-based GUI. The system inductively learns a landmark grammar, uses it to extract the marked-up fields from similar pages, and generates the

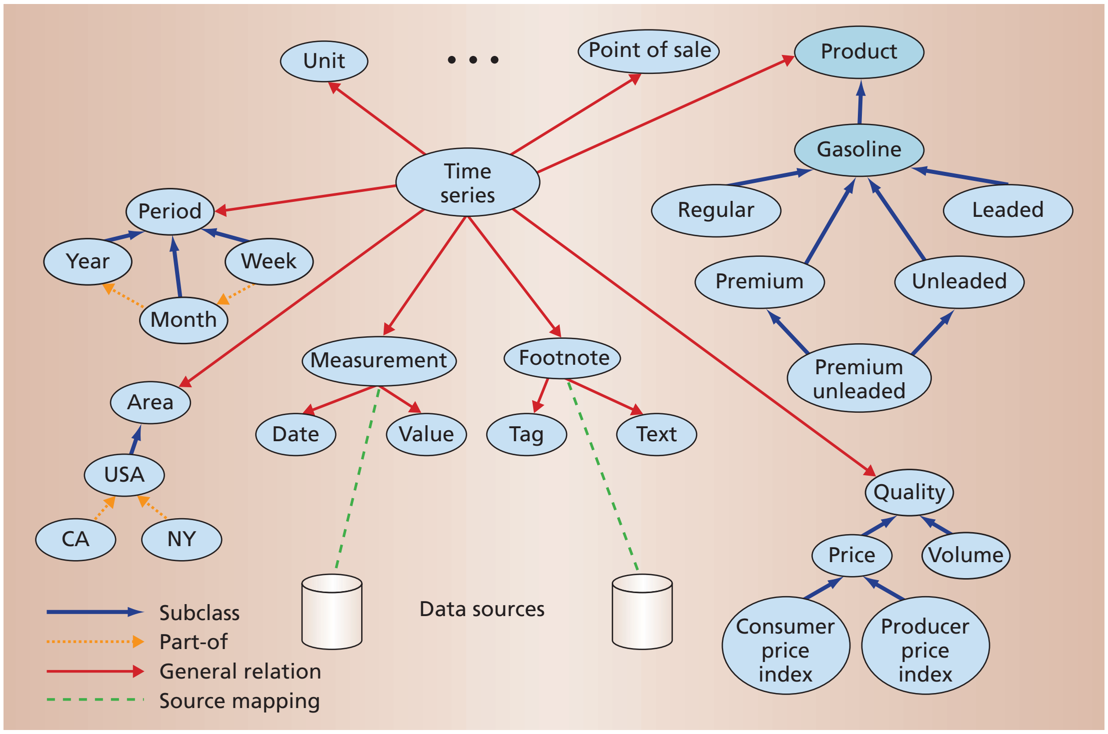

Figure 2. Fragment of an EDC domain model describing time-series data about different gasoline products. Each concept models a class of database contents. The user frames queries in terms of these concepts and their variants.

necessary wrapper code. The wrapper serves as a simple relational database that accepts parametrically defined SQL terms and dynamically retrieves data from the associated Web pages and forms.

To obtain uniformity and support reasoning, SIMS models each data source, whether natively relational or wrapped by Ariadne, by associating it with an appropriate domain-level concept description. A set of approximately 500 domain terms, organized in 10 subhierarchies, constitutes the EDC domain model thus far.

Figure 2 shows a fragment of an EDC domain model that describes time-series data about different gasoline products. Time-series data includes dimensions such as product, property measured (price, volume), location, and unit of measure. Each time series is defined by using a specific value for each of the hierarchical dimensions. For example, a source may provide the monthly prices (based on the consumer price index) of premium unleaded gasoline in California. The dimensions provide metadata that describes the series. A set of measurements—pairs of date and value—model the actual data.

The domain model also describes whether a source has footnotes for any of the data. If requested, the answer to a query will return footnote data associated with the corresponding tuples.

We linked these models into the Sensus ontology and added each of the retrievable time series and its 10 dimensional values to Sensus as a separate ontological concept. We also reified the relationships between series and dimensional values as Sensus relations—for example, has-product-type, area-of, and so on. This linking was semiautomatic. Using tools that facilitate the construction of wrappers and the semiautomatic description of sources is critical to cost-effectively scaling mediator systems to the large number of information sources available from government agencies.

Aggregation queries

We continue to investigate methods for integrating data sets and sources with information aggregated at different granularities and with different coverage. For example, one data source might include gasoline-price information for the entire US reported by month for the past 10 years; another source might have the information reported by year up to 1990; yet another source might contain yearly gasoline-price information discriminated by state. Our goal is to present a reasonably uniform view of available data without exposing this heterogeneity in aggregation granularity and coverage.

The main challenge of integrating data is dealing with data sets that exhibit varying granularity and coverage. For example, in one gasoline data set, the time attribute is month versus year, the geography is states versus countries, and the product is regular versus unleaded. In another gasoline data set, the time attribute is January 1978 through December 1986 versus January 1978 through January 1989; the geography attribute is San Diego, California, versus Boston, Massachusetts; and the product is leaded premium versus leaded regular.

Rather than mapping between domains or collecting metadata, we create mappings between the domain and an existing reference ontology.

Numerous statistical techniques, ranging from imputation to sophisticated forms of averaging, exist to deal with data on mismatched scales. The EDC's view of the data is sufficiently fine-grained to allow users to exploit most available information but coarse-grained enough to hide most granularity and coverage differences. We can usually correctly answer user queries with the available data sets. If the required data is unavailable, however, we have to reformulate a query that uses the available sources and combine these sources to provide an exact answer for that query.

In reformulated queries, data attributes such as time, geography, and product often follow natural granularity hierarchies—for example, day→month→year for time, city→state→country for geography. Combining the hierarchies results in a granularity *lattice*.[5] One node might therefore correspond to leaded gasoline data by month and state, another to leaded gasoline data by year and country.

We have developed algorithms to identify queries that we can answer exactly at each 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 it is about California and New York during the 1990-1999 time period. At runtime, we can then decide whether to answer a user query exactly or reformulate it using some distance function over the granularity lattice. To view initial results and algorithms for four BLS data sets for the average price of unleaded regular gasoline, visit http://db-pc01.cs.columbia.edu/digigov/Main.html.

Automatic footnote extraction

Footnotes are an important piece of metadata that often accompanies statistical tables. Footnotes can qualify the data of the entire table, a particular column, or specific cells in a table. Defining general procedures for the extraction of footnotes and determining the scope of their applicability is quite challenging when statistical tables come from text or HTML documents, as is the case with much available government data. Using finite-state analyzers that track each footnote's extent and associate footnote symbols with footnote text, we have automatically extracted footnotes as well as links between footnotes and text from Web pages and tables.

ONTOLOGY CONSTRUCTION

Practical experience demonstrates the difficulty of integrating different term sets and data definitions. Government-funded metadata initiatives have thus far yielded disappointing results. By focusing on collecting structural information such as formats, encoding, and links instead of content, these initiatives have generated large data collections of as many as 500,000 terms that are admirably neutral but unsuitable as "terminology brokers."

Rather than mapping between domains or collecting metadata, we create mappings between the domain and an existing *reference ontology*. This approach, tested on a relatively small scale in various applications, allows us to make any other domains mapped into the ontology available in the future to statistics agencies and eventually the general public. By making the reference ontology and our merging tools publicly available, we are encouraging others to align or even merge their term banks, data dictionaries, and so on.

Sensus ontology

We are collecting, aligning, and merging the contents of several large term banks and placing them under the high-level structure of the 90,000-node Sensus reference ontology.[6] Sensus is a large, fairly general ontology that links terms together into a subsumption (*is-a*) network, with additional links for part-of, pertains-to, and so on. This ontology is a rearrangement and extension of Princeton University's WordNet[7] that is retaxonomized under the Penman upper model, which was built at ISI to support natural language processing. Most of its content is identical to WordNet 1.5. Sensus is accessible via the Ontosaurus ontology browser at http://mozart.isi.edu:8003/sensus/sensus_frame.html.

Figure 3 shows the EDC project's ontology structure. To deploy Sensus, we first defined a domain model of approximately 500 nodes to represent the concepts in the EDC gasoline domain. We then used semiautomated alignment tools to link these domain concepts into Sensus. Finally, we defined a linguistically motivated connection we call *generally-associated-with* between the ontology and the domain model concepts, so the user can proceed rapidly from high-level concepts and related words to the domain model concepts that describe the database contents in detail. In turn, we created strict logical links from the domain model concepts into the databases themselves.

Semiautomated term alignment

The central problem with linking agency-specific domain models, as required by SIMS, to Sensus is determining where a given term belongs in such a large ontology. At first glance, aligning two ontologies or taxonomized term sets automatically might seem impossible. Most ontologies depend largely on non-machine-interpretable information such as concept names and English term definitions. However, recent research has uncovered a variety of heuristics that help with the identification and alignment process. We use a five-step, partially automated procedure consisting of

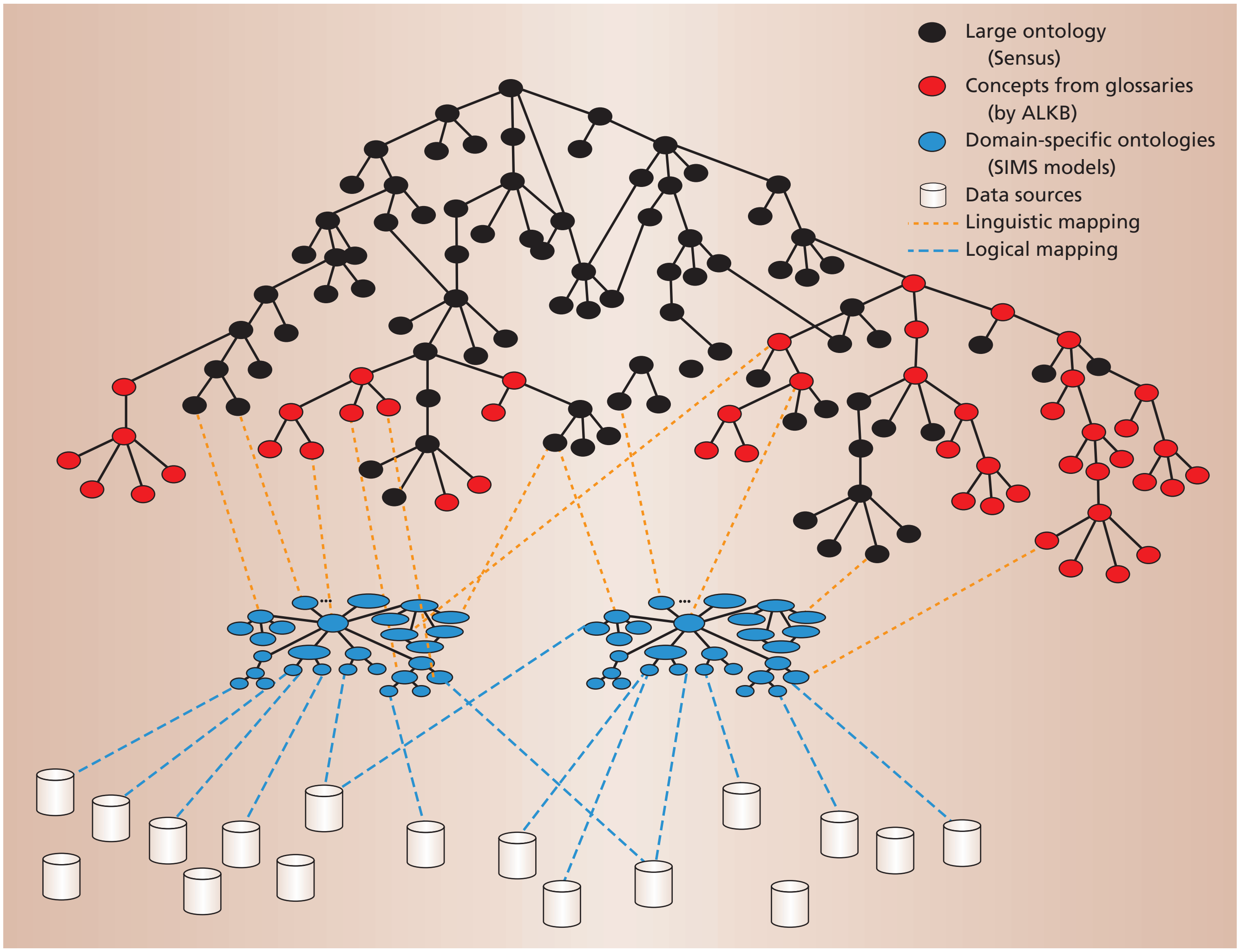

Figure 3. EDC ontology and domain models. New concepts derived from glossaries extended the Sensus ontology; then it was linked to the domain models of the actual gasoline data. In turn, the domain model nodes were linked to actual databases.

- heuristics that make initial cross-ontology alignment suggestions,
- a function for integrating these suggestions,
- alignment validation criteria and heuristics,
- a repeated integration cycle, and
- an evaluation metric.

The full power of these techniques for linking words from foreign lexicons[6] or concepts from other ontologies[8,9] is still under study. We have implemented two existing matching heuristics, NAME and DEFINITION MATCH, and developed a new one, DISPERSAL MATCH. NAME MATCH performs an exhaustive substring match of the concept name to be linked with every concept name in Sensus, with special rewards for beginning and ending substring overlaps. Because this match is very slow, we use an algorithm that matches gene sequences to obtain a two-order-of-magnitude speedup. After appropriate demorphing and closed-class word removal, DEFINITION MATCH compares the overlap of words in the definition of the concept to be linked against the definitions of all Sensus concepts. We have implemented a standard IR-based, vector-space-matching algorithm to improve efficiency.

DISPERSAL MATCH is based on the expectation that if a set of concepts to be linked are semantically related, they will tend to cluster together inside Sensus because Sensus concepts are also organized by semantic closeness. Applying this heuristic to link approximately 100 SIMS domain model concepts organized into 10 subgroups into Sensus provided almost-perfect accuracy in 8 of the 10 subgroups. We are currently using this method to link approximately 7,000 glossary items acquired from the EIA.

Extracting glossary entries

Several terminology problems require special attention in a cross-agency endeavor. The proliferation of terms and the different ways that agencies define ostensibly identical terms are particularly confusing to non-

Determining the expansion of an abbreviation or acronym is often domain dependent, making ambiguity one of the most challenging problems of acronyms.

specialists. For example, one database may use the term "wages," while another uses the term "salary" and also includes "wages" and "income." Glossaries are not always helpful because they often contain important information buried within lengthy definitions.

The first year's research on definition analysis had three objectives. The first was to identify a set of resources across agencies relevant to the energy-data domain. The second was to develop tools that automatically parse these definition sets, regardless of internal complexity. The final goal was to map the terms into a knowledge base with a uniform structure. We use this representation to map terms into Sensus, beginning semiautomatically but with increasing levels of automation as each component improves.

Columbia University's Automatic Lexical Knowledge Base (ALKB) (http://www.cs.columbia.edu/nlp/flkb/) uses a definition source such as a Web page or document to automatically create a structured system for use in generating ontologies and analyzing definitions. The basis of the system is

- prior experience using dictionary analysis to create a knowledge base,[10]
- automatic phrase variation rules for mapping related terms,[11]
- lexical resources for determining the range of ambiguity,
- verification in corpora to confirm ambiguity measures, and
- using linked phrases to eliminate potential ambiguity of single-word terms.

ALKB uses a combination of rule-based and statistical methods. After it applies a part-of-speech tagger to the definition, ALKB applies LinkIT, a noun-phrase chunker from Columbia, to determine linked noun phrases. It also counts occurrences of word pairs to find multiword collocations. Next, ALKB tags two types of semantic attributes. It determines predefined semantic attributes by analyzing the definition literature and a definition set. Predefined attributes such as "contains," "used for," "excludes," and "includes" are arranged in three categories: properties, excludes/includes, and quantifiers. ALKB also runs the bigram probability model across the entire document to find potential attributes that might help classify the document. It locates the assembled attributes in the definition it is currently analyzing and shows them to the user with the phrases surrounding them. This helps the user decide which attributes to add. We use the output from this analysis to build a framelike representation of each glossary entry to input into Sensus.

We have used ALKB to analyze nearly 7,000 definitions from several sources, including the EIA's Gasoline Glossary and its larger Glossary of Selected Terms and Abbreviations as well as about 35 relevant SIC and NAICS code metadata definitions and explanations from the Census Bureau. We also ran ALKB over medical definitions automatically extracted from lay articles in Columbia's Digital Library II project.[12] To resolve mapping issues associated with complex data, we are currently making ALKB output more automatically usable as Sensus input.

Acronym analysis

ALKB uses the Acrocat acronym cataloguing system to determine the meaning of acronyms. Acrocat lists confidence markers for each acronym in the current definition. We developed Acrocat as a subroutine of ALKB because agency-specific abbreviations and acronyms are often uninterpretable outside a given agency or domain. We built code for initial acronym resolution, and we are linking Acrocat with existing acronym and abbreviation glossaries to add guesses from these external resources.

Determining the expansion of an abbreviation or acronym is often domain dependent, making ambiguity one of the most challenging problems of acronyms. For example, NFS refers to Network File System in computer science, but it could mean Not For Sale in the auction world.

EDC INTERFACE

It is currently difficult to make productive use of the statistical data available on the Web. Because of the vast amount of information, existing systems typically offer two fundamentally different user interfaces. One access method exchanges generality for ease of use by relying on ready-made presentations consisting of tables and charts designed to answer typical questions. However, hundreds of these presentations may exist, making it hard to find the one that provides the closest answer. At best, these systems provide a keyword-searching mechanism to help users discover relevant presentations, but in many cases none of them address the user's specific query. The other method for finding information achieves generality by allowing users to construct their own queries. However, these interfaces are for experts only because they require intimate knowledge of the database's domain and structure, the meaning of the attributes, the query language, and ways of presenting the resulting information.

To address these access problems, we have developed a unified, Web-based user interface for querying and presenting statistical information. Our focus thus far has been on developing a robust, portable, and efficient interface that facilitates user access to data from multiple sources and 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 back-end integration of one or multiple individual databases, along with footnotes and links to original data sources; and
- ontology browsing that supports the entire integration model, displaying 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 with reference to the Sensus domain models. The design is obviously extensible: As we add new databases to the system and link their domain models into Sensus, their parameters are immediately available for querying. Sensus taxonomization ensures that the interface appropriately groups menu display options.

We have created a prototype of the kind of system required to support information access over heterogeneous databases developed and maintained by different government and private-sector agencies. Current EDC research focuses on three areas: the rapid inclusion of new databases into the system; the extraction and inclusion of additional, peripheral information related to data; and the development of sophisticated yet user-friendly interfaces tailored to the general public. During the next two years, we hope to demonstrate semiautomated database wrapping and modeling, restricted but free-form natural language query input in multiple languages, enhanced term extraction and glossary mining, caching and in-memory data processing for faster query response, and a more flexible user interface. ✷

Acknowledgments

This work was funded by the National Science Foundation's Digital Government Program under contract EIA-9876739. The authors are grateful for the hard work and dedication of four graduate students: Usha Ramachandran at USC/ISI and Jay Sandhaus, Anurag Singla, and Brian Whitman at Columbia University.

References

1. Y. Arens, C.A. Knoblock, and C-N. Hsu, "Query Processing in the SIMS Information Mediator," *Readings in Agents*, M.N. Huns and M.P. Singh, eds., Morgan Kaufmann, San Francisco, 1998, pp. 1005-1023.
2. J.L. Ambite and C.A. Knoblock, "Flexible and Scalable Cost-Based Query Planning in Mediators: A Transformational Approach," *Artificial Intelligence J.*, vol. 118, no. 1-2, 2000, pp. 115-161.
3. R. MacGregor, "The Evolving Technology of Classification-Based Knowledge Representation Systems," *Principles of Semantic Networks: Explorations in the Representation of Knowledge,* J. Sowa, ed., Morgan Kaufmann, San Francisco, 1990, pp. 82-101.
4. I. Muslea, S. Minton, and C.A. Knoblock, "Hierarchical Wrapper Induction for Semistructured Information Sources," *IEEE Trans. Knowledge and Data Eng.*, forthcoming.
5. V. Harinarayan, A. Rajaraman, and J.D. Ullman, "Implementing Data Cubes Efficiently," *Proc. ACM SIGMOD Conf.*, ACM Press, New York, 1996, pp. 15-27.
6. K. Knight and S.K. Luk, "Building a Large-Scale Knowledge Base for Machine Translation," *Proc. 12th AAAI Conf.*, MIT Press, Cambridge, Mass., 1994, pp. 773-778.
7. C. Fellbaum, ed., *WordNet: An On-Line Lexical Database and Some of Its Applications,* MIT Press, Cambridge, Mass., 1988.
8. G. Rigau and E. Agirre, "Disambiguating Bilingual Nominal Entries against WordNet," *Proc. 7th ESSLLI Symp.*, ESSLLI Organizers, Barcelona, Spain, 1995, pp. 15-19.
9. E.H. Hovy, "Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses," *Proc. 1st Int'l Conf. Language Resources and Evaluation (LREC),* European Language Resources Assoc., Paris, 1998, pp. 535-542.
10. J.L. Klavans, "COMPLEX: A Computational Lexicon for Natural Language Processing," *Proc. 12th Int'l Conf. Computational Linguistics (COLING 88),* Int'l Committee of Computational Linguistics, San Francisco, 1988, pp. 651-657.
11. J.L. Klavans, C. Jacquemin, and E. Tzoukermann, "A Natural Language Approach to Multi-Word Term Conflation," *Proc. DELOS Conf. European Research Consortium Information Management (ERCIM),* DELOS, Paris, 1997, pp. 153-161.
12. J.L. Klavans and S. Muresan, "DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text," *Proc. 2000 American Medical Informatics Assoc. (AMIA) Ann. Symp.*, AMIA, Bethesda, Md., 2000, pp. 56-64.

José Luis Ambite *is a research scientist at the University of Southern California's Information Sciences Institute. His research interests include database access planning. Ambite received a PhD in computer science from the University of Southern California. He is a member of the AAAI. Contact him at ambite@isi.edu.*

Yigal Arens*, co-principal investigator on the EDC project, is the director of the Intelligent Systems Division of the University of Southern California's Infor-*

mation Sciences Institute and co-director of the USC/Columbia University Digital Government Research Center. His research interests include information integration and planning in the domain of information servers (specifically heterogeneous databases), knowledge representation, and information-to-medium display planning in human-machine communication. Arens received a PhD in mathematics from the University of California, Berkeley. He is a member of the AAAI. Contact him at arens@ isi.edu.

***Eduard Hovy** is the director of the Natural Language Group at the University of Southern California's Information Sciences Institute and an associate research professor of computer science at USC and the University of Waterloo. His research interests include machine translation, automated text summarization, automated question answering, multilingual information retrieval, and the semiautomated construction of large lexicons and terminology banks. Hovy received a PhD from Yale University. He is president of the ACL and past president of the AMTA. Contact him at hovy@ isi.edu.*

***Andrew Philpot** is a scientific programmer at the University of Southern California's Information Sciences Institute. He received an MA in computer science from Stanford University. Contact him at philpot@isi.edu.*

***Luis Gravano** is an assistant professor of computer science at Columbia University. His research interests include data aggregation across heterogeneous databases and database access planning. Gravano received a PhD in computer science from Stanford University. Contact him at gravano@cs.columbia.edu.*

***Vasileios Hatzivassiloglou** is a research scientist in the Department of Computer Science at Columbia University. His research interests include using statistical techniques to address problems in natural-language text processing, including natural-language generation, multidocument text summarization, and term extraction from footnotes and other text. Hatzivassiloglou received a PhD from Columbia University. He is a member of the ACL. Contact him at vh@cs.columbia.edu.*

***Judith Klavans** is director of the Center for Research on Information Access at Columbia University. Her research interests include computational linguistics and natural-language processing, with an emphasis on digital libraries and lexicons. Klavans received a PhD in linguistics from the University of London. She is a member of the ACL. Contact her at klavans@cs. columbia. edu.*