A Portal for Access to Complex Distributed Information about Energy

Jose Luis Ambite[1], Yigal Arens[1], Walter Bourne[2], Peter T. Davis[2], Eduard H. Hovy[1], Judith L. Klavans[2], Andrew Philpot[1], Samuel Popper[2], Ken Ross[2], Ju-Ling Shih[2], Peter Sommer[2,] Surabhan (Nick) Temiyabutr[2], Laura Zadoff[2]

[1] Digital Government Research Center
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292

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

Abstract

The Digital Government Research Center (DGRC) has completed phase one of the Energy Data Collection (EDC) project. In this paper, we present the results of building and evaluating system components, along with plans for phase two of the project. Phase one focused on data about petroleum products' prices and volumes, provided by the Energy Information Administration, the Bureau of Labor Statistics, and the Census Bureau, and the California Energy Commission, in the form of over 50,000 data tables. This research centers on providing dynamically planned access to multiple non-homogeneous databases and other data collections, using a query planner and a large-scale (90,000-node) concept ontology and a domain model, both of which are accessed via various interfaces, including cascaded menus, a natural language question analyzer, and an ontology browser. Other data access research focuses on the caching and very fast display of massive amounts of data. In order to more rapidly construct the domain models, systems were developed for automatically identifying terminology glossary files from websites, extracting and formalizing the glossary definitions, clustering them appropriately, and automatically embedding them into the existing ontology and domain model. Work on evaluation focuses on measuring the effectiveness of the use of this system by a variety of users at various levels of expertise.

1. Introduction

Increasingly, federal, state and local Government need to make their information available to the online public. Frequently, though, there is too much data, it is in a variety of formats, and the data is only partly consistent internally (for example, term definitions created by different agencies differ; data values measured at different times seldom agree, etc.). In addition, the online user is seldom a domain expert, and hence requires an interface that allows him or her to browse, learn, and explore.

In an ideal world, some automated system would simply interpret all the Government data, integrate it into a single consistent view, and provide a simple-to-use interface for anyone, government expert and public alike, to find whatever they need and display it in whatever way they like. Many obstacles must be overcome for such a system to be built. The DGRC's[1] Energy Data Collection (EDC) project is addressing these obstacles. In this paper we briefly review the major problems and our approaches to solving them. We first describe the system architecture. Next we describe database access, database integration and homogenization using domain models and ontologies, and the semi-automated construction of domain models by glossary mining and term clustering. After discussing interfaces and evaluation, we conclude with an example of tech transfer and ongoing and future work.

---

[1] The Digital Government Research Center (DGRC; www.dgrc.org) was established in 1999. 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.

2. The EDC System

2.1 Architecture

The EDC project has been working on gasoline data with representatives of the Census Bureau, the Bureau of Labor Statistics (BLS), the Energy Information Administration (EIA) of the Department of Energy (DoE), and the California Energy Commission (CEC). For example, the EIA provides extensive monthly energy data to the public at http://www.eia.doe.gov. This site receives hundreds of thousands of hits a month, even though most of its information is available only as downloads of standard web (HTML) pages or as fixed pre-prepared PDF documents, and only for the last few years. The current facility thus supports only limited access to this very rich data source: it does not make visible the many definitions and footnotes that explain the complex nature of the data (whose changing definitions sometimes make incomparable figures appear to be comparable), and its query definition facility is too difficult for anyone but experts.

The EDC system (Ambite et al. 2001) (architecture in Figure 1) addresses both problems as follows:

Information Integration. We have investigated two different technologies for identifying, describing, and accessing the contents of databases; one focusing on live data in non-homogeneous formats (Section 2.2.1) and the other on fast access to cached data (Section 2.2.2).

Ontology Construction. We have extended USC/ISI's 70,000-node terminology taxonomy SENSUS to incorporate new energy-related domain models. We have developed techniques to help automate the creation of domain models by extracting terms from glossaries (Section 2.3.1), clustering them, and adding them to the ontology (Section 2.3.2).

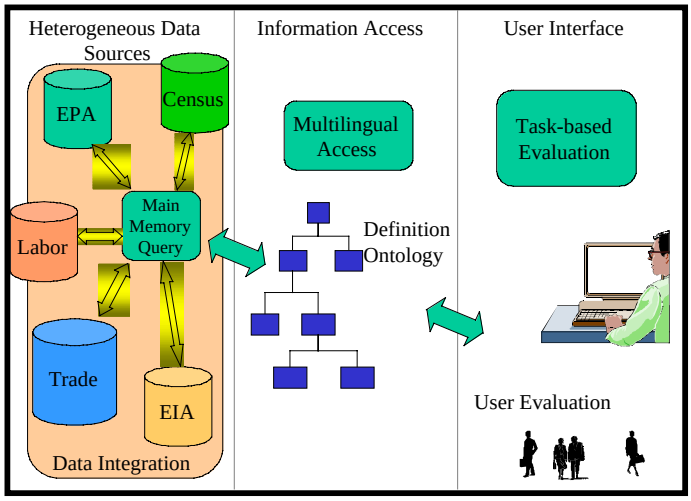

Figure 1. EDC system architecture; only partly realized.

User Interface Development. We have implemented various aspects of user-friendly interfaces, including query formation by ontology browsing (Section 2.4.1) and semi-free-form natural language input (Section 2.4.2). Its use is evaluated as described in Section 4.

2.2 Database Access and Processing

2.2.1 Access to Live Data in Non-Homogeneous Formats (ISI)

Retrieving data dispersed among multiple sources of various forms (HTML pages, databases, numerical tables in text format, etc.) requires familiarity with their contents and structure, query languages, and location. The access system must ultimately break down a retrieval task into a collection of specific queries to specific data sources. With a large number of sources, this can be a complex, time consuming problem.

Our approach to integrating statistical databases builds on research performed by the SIMS group at ISI (Arens et al. 1996). SIMS assumes that the system designer specifies a model (the *domain model*) of the application domain and defines the contents of each source (database, web server, etc.) in terms of this model. The SIMS planner provides a single point of access for all the information: the user expresses queries without needing to know anything about the individual sources. SIMS translates the user's high-level request, expressed in a subset of SQL, 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 are expressed internally in the Loom knowledge representation language (MacGregor 1990). The SQL subset is limited in its treatment of aggregation operators (such as sum, average, etc.).

Defining each data source in terms of the domain models, and creating its individual access routines, is known as *wrapping*. Using appropriate tools, we have wrapped 53,689 time series data tables in various formats, collected from the Energy Information Administration, the Census Bureau, the Bureau of Labor Statistics, and the California Energy Commission:

EIA/OGIRS databases (Oracle and Microsoft Access)	29,542
EIA/PSM web pages, text files, PDF files	24,181
BLS web pages (HTML)	53
CEC web pages (HTML)	3

See (Hovy et al. 2001; Ambite et al. 2002a), or visit http://www.dgrc.org/, click on *Research* and then on *related pages and data*.

2.2.2 Fast Data Access to Cached Data (Columbia)

Professor Ross and his group have been working on efficient query processing. Within the scope of the digital government project, there is particular interest in rapidly computing aggregates of large numbers (millions) of records. The goal is to enable the interactive exploration of data sets such as the PUMS data from the Census Bureau via some form of dynamic query interface.

The technical means for achieving fast query response are centered around storing large datasets in main memory. Now that multi-gigabyte main-memories are affordable, it is clearly feasible to store large datasets in RAM for analysis. In that context, the design of in-memory index structures needs to take issues such as CPU cache miss latencies into account (Rao and Ross 1999; 2000; Ross et al. 2001).

One query processing technique that has been recently developed involves optimizing the way a query is executed to minimize branch misprediction penalties (Ross 2002). Branch misprediction latency can be a significant component of the response time of certain kinds of queries. A second technique involves using SIMD parallelism available on commodity processors to speed up database operations (Zhou and Ross 2002).

2.3 Automatic Ontology and Domain Model Building

In order to facilitate cross-domain integration, the domain models described in Section 2.1.1 are embedded into a single all-encompassing general model called the *ontology*. We have reimplemented, updated, and extended ISI's SENSUS ontology of 70,000 nodes (Knight and Luk 1994; Hovy et al. 2001). We have also developed techniques to facilitate domain model construction by extracting likely modeling terms from glossaries and other natural language text (Section 2.3.1) and then clustering them and embedding them into the existing ontology (Section 2.3.2).

2.3.1 Terminology Search and Formalization (Columbia)

We have developed an end-to-end system, GlossIT, for automatically identifying glossaries within large government websites, and then processing the entries in these glossaries to extract core ontological information. The output of GlossIT is then ready to be loaded into a large ontology for user access across agencies, domains, and databases (see Section 2.3.2). The first module, GetGloss, crawls a given domain and uses a rule-based system to decide if a given page or set of pages constitutes a glossary. The second module, ParseGloss, takes the output of GetGloss and first defines the glossary type (e.g. encyclopedic, acronyms, etc.) Given the glossary type, ParseGloss then analyzes definitions and creates a data structure for loading into the ontology. Fuller details are presented in (Klavans et al. 2002) and in www.cs.columbia.edu/digigov.

At the core of GetGloss is a rule-based algorithm that analyzes a large set of HTML pages, and produces a set of pages determined to be glossaries. GetGloss takes an input list of sites to crawl and crawls each one to a recursion depth of five to seven.. GetGloss then examines each page produced by the crawler. If a certain page looks structurally like a glossary, this page will then be added to the set of glossary pages. This algorithm requires no training, instead attempting to overcome the wide formatting differences, as well as the possibility of tags not having a true structural meaning. It does this through analyzing page structure, and looking for tags that most likely give some sort of clue towards the structure of the content within the page.

ParseGloss is designed to extract attributes and relationships from glossary entries. In its first step of parsing a glossary, ParseGloss uses a part-of-speech tagger on a definition to associate each word with its proper part of speech. It then sends the marked-up glossary entry to a noun-phrase chunker. The noun phrases are used to find the genus phrase, and head genus term of the glossary entry. The genus phrase is usually the first noun phrase in the

definition, and the head genus word is the head noun in the genus phrase. After finding the genus phrase and head genus word, ParseGloss uses a system of rules to extract properties from the glossary entries. In addition, it calculates the bigrams in the glossary, and uses the AcroCat system (Klavans and Whitman 2001) to obtain expansions for acronyms contained in the text.

2.3.2 Clustering and Ontology Merging (ISI)

We have implemented and tested fast versions of several term clustering algorithms, including SLINK, CLINK, and k-Nearest Neighbor. In order to merge clusters and/or individual terms into SENSUS, we have extended the matching heuristics described in (Hovy 1998) (NAME and DEFINITION MATCH) and developed a new one (DISPERSAL MATCH) (Hovy et al. 2001). In a small test experiment, we manually aligned 42 terms from the EIAGOE2 (EIA Glossary of Energy Terms) ontology as extracted by GetGloss to the SIMS domain model. Twenty-two of the EIA terms matched one or more SIMS model terms fully or partially; partial matches were annotated with an estimated confidence. Twenty of the terms were deemed not to correspond to any term at all. We then applied 12 variants of the NAME MATCH algorithm to the two ontologies, specifying the return of up to 15 candidate alignments, ranked by computed similarity. Details about to what degree the alignment algorithms discovered the manually specified alignment are shown in http://edc.isi.edu/alignment/.

A fairly extensive series of experiments focused on determining the optimal parameter settings for these algorithms, using three sets of data: the abovementioned EDC gasoline data, the NHANES collection of 20,000 rows of 1238 fields from a survey by the National Center for Health Statistics, and (for control purposes) a set of 60 concepts in 3 clusters extracted from SENSUS.

2.4 User Interfaces—Query Entry and Browsing

2.4.1 Ontology Browsing and Query Formation Interfaces (Columbia)

Our user interface work has concentrated on the development of fisheye/nonlinear magnification techniques to present large amounts of data in a relatively small space. Over the year, we have progressed from approximating a physical model of magnification to designing a more abstract one, maintaining the concept of emphasizing the fisheye foci (points of interest) within an overview of the surrounding data.

One approach that we have been exploring relies on a dynamic representation of unallocated display space (Bell and Feiner 2000) to position and scale data objects near the foci. This avoids overlap, while minimizing the need to scale down objects farther from the foci. The display space representation is also used to determine where to display detailed information about selected data objects that we obtain from the ontology. Rather than presenting this information in a separate window or in a fixed location in the current window, we instead display it in a box whose location is automatically chosen to be as close as possible to the selected data object without overlapping it or any of the other data objects. We have also been involved in user evaluations of these techniques, meeting with a user evaluation group.

2.4.2 NL Input Parsing for Query Formation (ISI)

In order to support query formation by non-expert users, we have implemented a semi-free-form natural language parser. AskCal (Philpot et al. 2002) combines language based techniques, including ATN parsing of free-form text queries, user modification of predefined template queries, and fall-back parsing by picking out landmark terms, to support a wide variety of user queries while reducing user query formulation effort. The system employs ontology- and data-based feedback mechanisms to guide the user toward regions of the query space where useful data can be found.

3. Evaluation (Columbia)

We conducted a formative evaluation of the DGRC database interface as an illuminative study. Our goal was to optimize interface effectiveness by identifying user needs and usability problems. The aim of this first stage of the evaluation was to inform design decisions by providing feedback and advice for the development of the interface.

We designed and conducted several methods in this first phase of the evaluation in accordance to the work of the interface development team. Our techniques included: 1) Development of heuristics for graphical user interfaces for databases; 2) Contrast analysis between conventional form-based menu systems and fisheye technology enabled interfaces; 3) Interviews with data and user experts as part of the research on database user needs, behaviors,

preferences, reactions, and perceptions to database interface design; 4) Resource analysis on inherent characteristics and challenges of both energy and census data; 5) Component analysis of applications using fisheye technology, so as to consider interface possibilities as innovative solutions.

In order to inform the design about the needs and searching characteristics of the intended audiences, we constructed a user type categorization based on levels and areas of expertise and corresponding purposes and tasks. Our database query log analysis on the EDS DataGate allowed us to configure characteristics that are common in database users and provided important information for the refining of the heuristics.

Major issues of the use of census data can be generalized to other databases that comprise massive amount of variables structured in complex hierarchies. The challenges are manifold. They include users' difficulties in defining their queries, and visualizing the context and the structure of information. Users also have difficulties in utilizing alternative terminologies in their queries, and in understanding variables that lack consistency over time and types of census database. Learning to use a non-traditional interface, however, appears to be an initial obstacle to most users.

The component analysis is the exploration of the specific elements of the fisheye technology enabled interface, namely, item magnification with full context visualization, searchlight overlays that provide just-in-time information, and manipulative enlargement of selected content. Potential solutions to the problems related to the data include using dynamic menus to incorporate oversized categories, and using transparent layers to present terminology definitions, alternative terms, ontological relationships, possible number of query results, annotations associated with data limitations and modifications, as well as cross-references with resources in other databases. These functions will assist users to define and refine queries, and retrieve information via multiple pathways from which both novice and expert users can benefit.

Our evaluation plan in its second phase involves the design and execution of user and task analysis. We will design task scenarios targeted to the defined types of user groups in which each participant will accomplish a realistic goal using the database interface prototype. The task will be completed with purposeful observation and followed by interviews to explore internal motivations and cognition. Usability testing will be the next step, to be followed by a summative evaluation with of the fully functioning interface.

4. Technology Transfer (ISI)

Wrapping the EIA's gasoline webpages led to an unexpected opportunity for technology transfer. The EIA's site http://www.eia.gov contains pages in which data from various states has been combined. The EIA, however, was eager to break out the data by state and recombine it. As shown on http://altamira.isi.edu/textwrap/, we developed an 8-step procedure that splits each page into head, body, and foot sections, extracts data using wrappers into a neutral single page CSV format file, and publishes the resulting relations in per-state HTML breakouts that mimic the original report style. Extracted data includes footnotes, which must be re-numbered as appropriate.

The EIA funded the construction of a general tool to perform this and similar operations on new files. We constructed the architecture and interface, and collaborated with Fetch Inc., a wrapper creation company, who built the wrappers. The tool was completed and delivered to the EIA in May 2002.

5. Moving to New Data (Columbia and ISI)

Ongoing work involves developments on two fronts: extending the system's current capabilities and developing new application areas.

We are extending the system's capabilities is several directions. With respect to the user interface, we will develop new interfaces to support additional interaction paradigms. We will also extend the natural language input interpretation (Section 2.4.2) to handle foreign languages, including Spanish (and possibly Mandarin Chinese). This work involves not only implementing a Spanish grammar and lexicon but also adding Spanish concept names to the Ontology, and possibly adding Spanish glossary definitions.

We are in the process of building an evaluation corpus to test our algorithms. The evaluation team (see Section 3) is designing further experiments with users to test the value of ontologies for browsing and access to government data.

A challenging new domain for applying and extending our techniques for information integration is transportation planning. In particular, we are developing a novel approach to commodity flow estimation in metropolitan areas in collaboration with faculty from the School of Planning, Policy, and Development of the University of Southern California (Ambite et al. 2002b).

Current approaches for freight flow estimation are no longer adequate given the growing complexity of transportation flows. Origin-destination surveys are increasingly expensive. Their continuous updating is more important than ever and adds to the expense of relying on such surveys. Thus, we propose to estimate freight flows from secondary data sources (Gordon and Pan 2001).

From the computer science research perspective, the transportation-planning domain presents several novel challenges. First, the domain requires integrating geospatial data sources, such as census tracts, traffic analysis zones, road networks, etc, in addition to traditional databases and web sources. Second, the available information sources only provide secondary data, so our system must derive new data and integrate it with other refined information in order to obtain sources and destinations of freight flows. Finally, the flow data must be assigned to the highway network, which involves applying complex network algorithms. In summary, our approach must address both integration of heterogeneous data and complex analysis of this data seamlessly. We are currently developing techniques for geospatial data integration and to manage the computation workflow for the derived data.

References

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.

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).

Ambite, J.L., Y. Arens, E.H. Hovy, A. Philpot, L. Gravano, V. Hatzivassiloglou, J.L. Klavans. 2001. Simplifying Data Access: The Energy Data Collection Project. *IEEE Computer* 34(2).

Ambite, J.L., Y. Arens, L. Gravano, V. Hatzivassiloglou, E.H. Hovy, J.L. Klavans, A. Philpot, U. Ramachandran, K. Ross, J. Sandhaus, D. Sarioz, A. Singla, and B. Whitman. 2002a. Data Integration and Access: The Digital Government Research Center's Energy Data Collection (EDC) Project In W. Mcver (ed), Kluwer Academic Publishers, to appear.

Ambite, J.L., G. Giuliano, P. Gordon, Q. Pan, and S. Bhattacharjee. 2002b. Integrating heterogeneous data sources for better freight flow analysis and planning. *Proceedings of the NSF's National Conference on Digital Government dg.o 2002.* Los Angeles, CA.

Bell, B. and S. Feiner. 2000. Dynamic Space Management for User Interfaces. *Proceedings of the ACM UIST Symposium. on User Interface Software and Technology, CHI Letters* 2(2): 239–248. San Diego, CA.

Gordon, P. and Q. Pan. 2001. Assembling and Processing Freight Shipment Data: Developing a GIS-Based Origin-Destination Matrix for Southern California Freight Flows. Technical Report.,National Center for Metropolitan Transportation Research (www.metrans.org).

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., A. Philpot, J.L. Ambite, Y. Arens, J. Klavans, W. Bourne, and D. Saroz. 2001. Data Acquisition and Integration in the DGRC's Energy Data Collection Project. *Proceedings of the NSF's National Conference on Digital Government dg.o 2001.* Los Angeles.

Klavans, J.L. and B. Whitman. 2001. ***.

Klavans, J.L., S. Popper, and ***. 2002. ***. *Proceedings of the NSF's National Conference on Digital Government dg.o 2002.* Los Angeles, 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.

Philpot, A., J.L. Ambite, E.H. Hovy. 2002. DGRC AskCal: Natural Language Question Answering for Energy Time Series. *Proceedings of the NSF's National Conference on Digital Government dg.o 2002.* Los Angeles.

Rao, J. and K.A. Ross. 2000. Making B+-Trees Cache Conscious in Main Memory. *Proceedings of the 2000 SIGMOD Conference*.

Rao , J.and K.A. Ross. 1999. Cache Conscious Indexing for Decision-Support in Main Memory. *Proceedings of the 1999 VLDB Conference*.

Ross, K.A. 2002. Conjunctive Selection Conditions in Main Memory. *Proceedings of the 2002 PODS Conference.*

Ross, K.A., I. Sitzmann, and P.J. Stuckey. 2001. Cost-Based Unbalanced R-Trees. *Proceedings of the 2001 SSDBM Conference.*

Zhou, J. and K.A. Ross. 2002. Implementing Database Operations Using SIMD Instructions. *Proceedings of the 2002 SIGMOD Conference*.