

Scalable Access and Integration of Statistical Data for Digital Government

Jose Luis Ambite*, Yigal Arens*, Eduard Hovy*, Judith Klavans & Andrew Philpot*

Digital Government Research Center

* Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
310-822-1511
{ambite,arens,hovy,philpot}@isi.edu

Department of Computer Science
Columbia University
535 West 114th Street, MC 1103
New York, NY 10027
212-854-7443
klavans@cs.columbia.edu

Abstract

The massive amount of statistical and text data available from government agencies has created a set of daunting challenges to both the research and analysis communities. These problems include heterogeneity, size, distribution, and control of terminology. At the Digital Government Research Center (www.dgrc.org) we are investigating solutions to these key problems. In this paper we focus on scalability of data integration across multiple databases and web sources, and ontology construction and mapping for terminology standardization. This collaboration between researchers from the Information Sciences Institute of the University of Southern California and the Department of Computer Science of Columbia University employs technology developed at both locations, in particular the SIMS multi-database access planner [AKH96,AK00], the SENSUS ontology [KL94,SPKR96,H98] and the LEXING automated dictionary and terminology analysis system [KM00,KW01]. Our application targets gasoline data from the Bureau of Labor Statistics, the Energy Information Administration of the Department of Energy, the Census Bureau, and other government agencies (see [AAPH+01] for an overview of the project).

Key words —Information Integration, Heterogeneous Data Sources, Web Wrappers, Ontology construction, Efficient Query Processing, Data Warehousing

1. Introduction: The Digital Government Research Center

The Government has a mandate to make its information available to the public. As access to the web becomes a household commodity, citizens will expect more and better quality data available to them from government agencies, in particular, from statistical federal agencies such as the Census Bureau, the Bureau of Labor Statistics, the Energy Information Administration, and others. 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.

The Digital Government Research Center (DGRC; www.dgrc.org) 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.

In this paper we describe the general framework for scalable information integration and access that we are developing at the DGRC for the Energy Data Collection project.



2. The Energy Data Collection Project

The DGRC Energy Data Collection (EDC) Project is being developed under the auspices of the National Science Foundation's Digital Government program. In this project we are working with representatives of Federal and State statistics agencies and other organizations to build a system for disseminating statistical data from 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). The EIA's web site, <http://www.eia.doe.gov>, is a representative example of the types of information with which we are dealing. This site provides extensive up-to-date energy data to the public, receiving hundreds of thousands of hits a month. However, most of its information is available only as downloads of standard web (HTML) pages or as prepared text or PDF documents. The current facility thus supports only a 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 a database-like query capability is not available. Moreover, we want to provide access not only to the EIA web site but to a host of other energy-related data sources.

The EDC project addresses these unique challenges building on our work on information integration [AKH96,AK00], automatic wrapper learning [MMK98], text analysis [K88,KM00,KW01], and ontology construction [KL94] and alignment [ACRR+94, OH94, H98].

Information integration. We have developed effective methods to describe and integrate the contents of multiple heterogeneous databases and web sources, allowing the user to accurately and efficiently query useful information while being insulated from the distribution of the data, the structure of the sources, their languages, formats and dependencies among the data. Currently, our system provides access to more than 50,000 time series on energy-related data integrated from databases and web sources of several federal agencies. We have developed *wrappers* to provide database-like access to information that is only available in text and HTML formats. These wrappers were efficiently constructed in a semi-automated fashion [MMK98]. Our system allows not only query access to the data but also provides footnotes and other types of metadata.

Ontology construction and alignment. In order to provide an intuitive access to the information, and facilitate the inclusion of new sources and domains into the system, all our models are linked to an over-arching ontology called SENSUS. SENSUS is a large taxonomy that includes over 90,000 common-sense concepts (senses). We have extended SENSUS with new energy-related concepts, and we have developed automated, concept-to-ontology alignment algorithms. We have created a cross-agency ontology that includes 7000 terms automatically extracted from multiple text glossaries available in documentation of the agencies. Moreover, this new terminology was semi-automatically linked to the SENSUS ontology using several alignment heuristics (such as name match, definition match, and dispersal match).

User interface. We have designed and implemented a flexible user interface to facilitate query construction and presentation of results. The interface offers several access modes to the information. First, it provides an ontology browsing mode in which a non-expert user can explore the application domain starting from general English terms and is progressively guided towards the energy-related concepts used in our domain to finally identify the time series data of interest. For non-expert users we also plan to incorporate natural-language searching mechanisms. A second mode is a database query interface for more advanced users. The user specifies some of the metadata about the time series of interest regardless of their location or format. The integration system transparently retrieves the required information.

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 system architecture and the phases of the lifecycle of the system are shown in Figure 1. The ontology construction phase includes the work on semi-automated term alignment, term extraction from glossaries, and acronym handling. During the user phase, the interface facilitates the construction of queries by the user, which may involve ontology browsing and other interaction methods. The user interface dispatches a high-level query to the query processor, which, in turn, returns the results to the interface for appropriate display. Finally, in the access phase the query planner consults the source descriptions in the ontology and transforms the high-level user query in an optimized query plan that accesses the relevant sources, retrieves and composes the requested information.

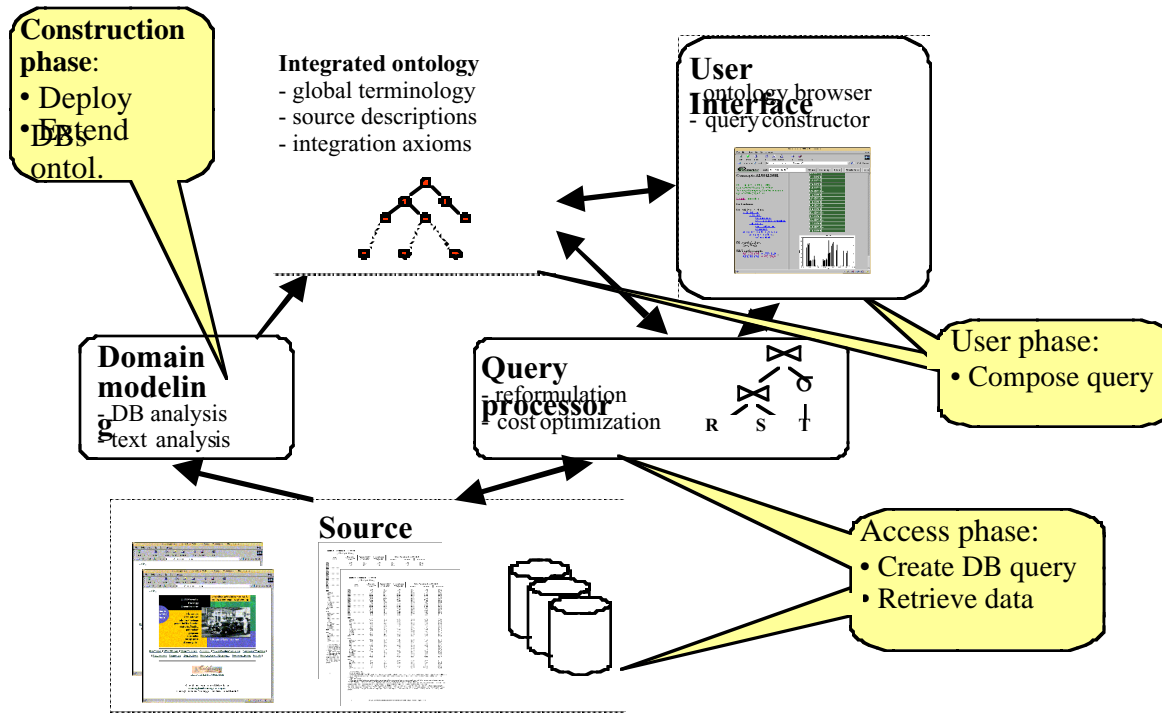


Fig. 1. Architecture of the EDC System.

3. Information Integration

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

3.1 Domain Models

Our approach to integrating statistical databases builds on research performed by the SIMS group at ISI [AKH96,AK00]. SIMS assumes that the system designer specifies a global model of the application domain and describes the contents of each source (database, web server, etc.) in terms of this global model. A SIMS mediator 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* [AK00], 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 [M90].

In SIMS, each of these data sources is modeled by associating it to an appropriate domain-level concept description. A set of approximately 500 domain terms, organized in 10 sub-hierarchies, 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., 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, the time series EIA-T31-1 is described as providing the monthly prices of regular gasoline sold to end-users in the state of Maine measured in cents per

[data_publications/petroleum_marketing_monthly/current/txt/tables31.txt](#)). The resulting time-series were incorporated to our system. More details in this wrapping effort can be found in [HPAA+01].

- **BLS**: 53 time series, all of which are obtained by a wrapper which instantiates a form on the BLS web site: 27 BLS/WP PPI Revision Commodities , 12 BLS/PP PPI Revision Current Series , 8 BLS/AP CPI Average Price Data , and 6 BLS/CU CPI All Urban Consumers.
- **CEC**: 3 CEC series, obtained by wrapping a single page on the California Energy Commission web site.
- **EIA PSM**: 16 EIA PSM (Petroleum Supply Monthly) series. PSM is a period publication. We have converted a snapshot of the PSM in pdf format to html, then wrapped it to obtain the 8 annual and 8 monthly measurements.
- **EIA OGIRS**: 25,000 modeled EIA OGIRS (Oil and Gas Information Resource System) series. OGIRS is a data product available from EIA on CDROM; it is implemented as a query system on top of MS Access. We can access the Access data directly; also, we have imported the data themselves into Oracle. OGIRS has much formal metadata that can be viewed in Access. We captured the metadata in our model and provided uniform access to all these series.

All 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 the next section, 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.

4. Ontology Construction

We would like to help automate the process of creating Domain Models. Our approach is to use machine learning and data mining techniques to identify important characteristics of the data source in order to be able to recognize what kind(s) of data are included. This very ambitious (and hence long-term) goal involves breaking down the work as follows. Given some new domain:

- Text (glossaries, manuals, etc. associated with the data): extract information from the text, using Finite State and statistical techniques [KM00]; create formally defined concepts
- Databases: create concepts out of the metadata
- Domain Model: use this information to build a domain model
- Ontology matching: cluster and embed these concepts into SENSUS [H98].

4.1 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 metadata initiatives with rather disappointing results. The focus has been on collecting structural information (formats, encodings, links), instead of content, resulting in large data collections (up to 500,000 terms) that are admirably neutral, but unsuitable as terminology brokers .

We are following a different approach, one that has been tested, on a relatively small scale, in various applications in the past two years. Rather than mapping between domains or collecting metadata, we create mappings between the domain and an existing *reference ontology*. This choice allows us in the future to make available to statistics agencies (and eventually to the general public) any other domains that have also been mapped into the reference ontology. Furthermore, by making publicly available the reference ontology with our merging tools, we hope to encourage others to align (or even to merge) their termbanks, data dictionaries, etc., as well.

We are collecting, aligning, and merging the contents of several large termbanks, placing them under the high-level structure of an existing large (90,000-node) and fairly general ontology called SENSUS, built at USC/ISI [KL94]. Its terms are 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 [F98] (built at Princeton University on general cognitive principles), retaxonomized under the Penman Upper Model [BKMW89] (built at ISI to support natural language processing). For most of its content, SENSUS is identical to WordNet, SENSUS can be accessed using Ontosaurus, the ontology browser at <http://mozart.isi.edu:8003/sensus2/> [SPKR96].

The ontology for the EDC project has the structure shown in Figure 3. We added to SENSUS the 500 concepts in the SIMS domain model, the 50000 concepts that describe the source time-series EDC gasoline domain, the 7000 concepts automatically extracted from glossaries. We linked these domain concepts into SENSUS using semi-automated alignment tools. This linking allows the user to browse rapidly from high-level concepts to the concepts associated with real data in the databases.

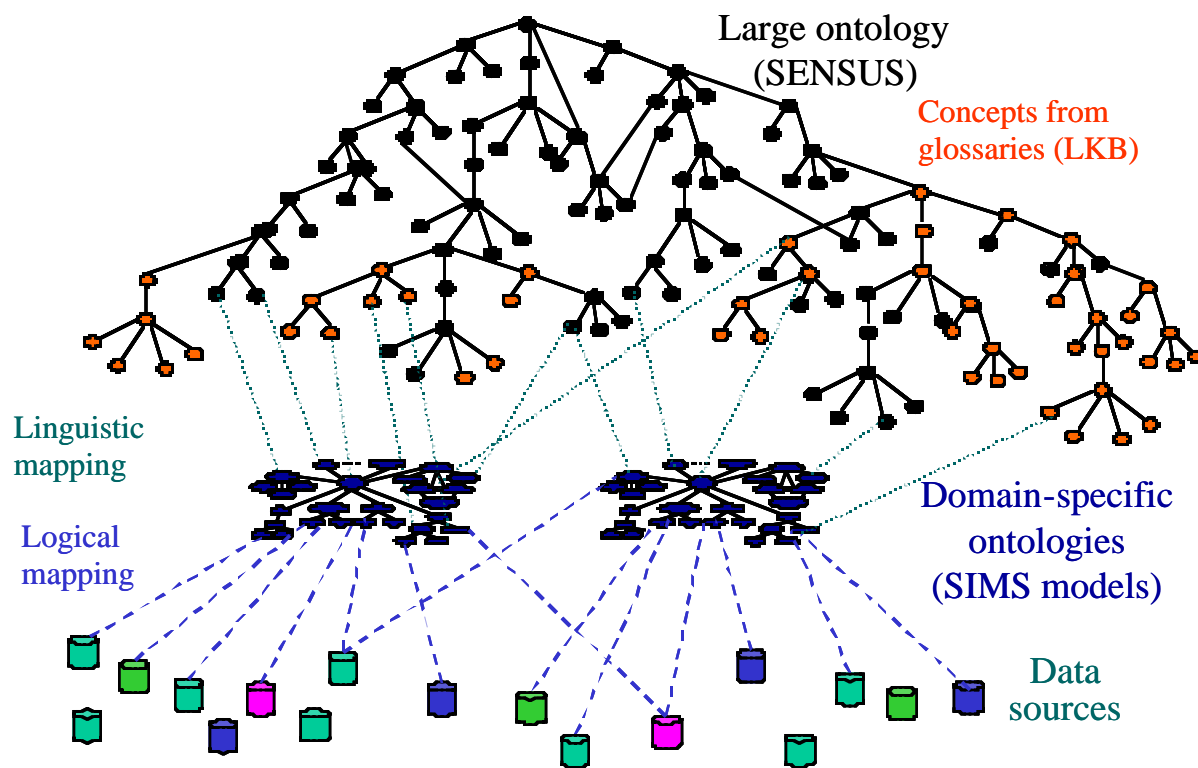


Fig. 3. Ontology, Domain Models, and Data Sources

4.2 Extracting Concept Definitions from Text Glossaries

Several terminology problems require special attention in a cross-agency endeavor. In particular, and especially confusing to non-specialist users, is the proliferation of terms, and the fact that agencies define ostensibly identical terms differently. What is called *wages* in one database may be what another calls *salary* (even though it may have a *wages* too, as well as an *income*). However, many agencies offer glossaries that define in natural language text the various terms. An example of a glossary definition in our gasoline domain is shown in Figure 4. Analyzing these definitions provides valuable insights when modeling an application domain. Automating this analysis is crucial for scalability.

Motor Gasoline Blending Components: Naphthas (e.g., straight-run gasoline, alkylate, reformate, benzene, toluene, xylene) used for blending or compounding into finished motor gasoline. These components include reformulated gasoline blendstock f or *{sic}* oxygenate blending (RBOB) but exclude oxygenates (alcohols, ethers), butane, and pentanes plus. *Note:* Oxygenates are reported as individual components and are included in the total for other hydrocarbons, hydrogens, and oxygenates.

Fig. 4. Sample Glossary Definition

We have built a system which extracts the genus word and phrase from free-form definition text, entitled LEXING, for **Lexical Information from Glossaries**. The extractions were used to build automatically a lexical knowledge base from on-line domain specific glossary sources. We combine statistical and semantic processes to extract these terms, and demonstrate that this combination allows us to predict the genus even in difficult situations such as empty head definitions or verb definitions. We also employ the use of linking prepositions for use in skipping past empty head genus phrases [KM00,KW01,K88].

Genus Term and Phrase Finding. The method LEXING uses to determine the genus uses knowledge gained from the part-of-speech tagging and noun-phrase (NP) chunking components. We have developed a grammar of phrase identification from a manual study of various definitional sources, and have implemented an evaluation metric for comparing our system's results against a manually tagged set of 500 glossary definitions from 5 different sources.

Definition	(Head Term:)	(Definition Text)
Definition Text	(Genus Phrase (GP))	(Remainder)
Remainder	Text	
Genus Phrase	NP (of)?	(Genus Phrase)
Genus Term	(last noun of first GP NP)	

Figure 6 - Genus Phrase and Term Grammar

Since each domain could use domain specific semantic separators, we also introduced the notion of *automatically derived semantic attributes* that are inferred simply from their frequency in the text. LEXING identifies separators such as *having a*, *used for*, or *containing a*, as cue phrases suitable for semantic chunking. As an example, the phrase *having a* was not in our original list of manually-derived separators, but after running a bigram analysis, we discovered its frequency and importance. Below we show an abbreviated parse of our sample definition, showing the semantic separators *used-for* and *excludes* as well as the genus term (Naphthas) and acronym.

<p>(term: Motor Gasoline Blending Components (full-def:) (core-def:) (is-a Naphthas) (properties (used-for blending) (excludes oxygenates)) (acronym RBOB))</p>

Figure 7 – Sample Glossary Parse with LEXING

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 is listed with confidence markers. Acrocat was developed 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. We have built code for initial acronym resolution and are linking Acrocat with existing acronym and abbreviation glossaries in order to add guesses from these external resources.

4.3 Concept Clustering and Alignment

Having extracted approx. 7000 concepts from glossaries and other text acquired from EIA and BLS, our challenge was to link these concepts into SENSUS at their most appropriate points. We developed the process shown in Figure 8, with times required to handle the 7000 glossary concepts (this work is almost complete, but still needs evaluation):

1. Prepare the concepts to be linked: format them appropriately, stem definitions, etc.

2. Cluster the concepts into semantically related clusters of between 20 and 50 concepts each. For this we tested various clustering algorithms and implemented a fast version of k-Means.
3. Per cluster, apply the Name and Definition Match algorithms [H98] to each concept, and collect for each concept the candidate matches in SENSUS (see below).
4. Apply the Dispersal Match algorithm (newly invented) to each cluster, in order to select from each set of candidate matches the most appropriate one for each concept. Return for each concept a ranked list of alignments.
5. Evaluate the alignments manually, and improve them as required.

Clustering: For a given set of concepts, cluster them to find the sets of ones that belong together semantically. One of three criteria has to be provided by the user in determining cluster cutoff:

- number of clusters desired
- similarity cutoff value
- max/min size of desired clusters
- In addition, the user has to specify how the inter-subcluster distance is measured; variations include
- slink, clink, median, and Ward's method
- speedup: recursive nearest-neighbor (RNN)
- usually using a Euclidean distance (similarity) metric, defined in a vector space of word tokens.

We tried the above clustering methods (using ISI's clustering package ISICL), after having reimplemented some of them to work more efficiently and to link to a GUI that would display the results graphically. We then implemented a version of k-Means clustering for text and numeric vectors, with both Euclidean and spherical distance measures.

Name Match: Match the name of the concept against the name of each ontology concept and ranks the goodness of fit. Variations:

- longest common substring (contiguous)
- position of match: equal beginnings, ends, or wholly contained
- prefix/suffix structure: *leaded vs unleaded, happy vs happiness*
- stemming: root words or inflected words
- non-word matches: letter trigram match
- case sensitivity

Name match includes a substring match operation, which in our first implementation was extremely slow. We later implemented one version of an algorithm used in molecular biology to match DNA subsequences. The results were still daunting over 150 hours to match 6,000 concept names against all of SENSUS. For this reason we implemented a match of just letter trigrams (breaking the name into all its contiguous 3-letter units), which took approx. 24 hours but gave less satisfactory results.

Definition Match: Match the English definition of the concept against the English definition of each ontology concept and rank the goodness of fit. Variations:

- text to match: word tokens or letter trigrams
- stemming: root forms or fully inflected
- include or exclude stop words
- terms to match: basic definition vs. expanded definitions (SENSUS synonym sets)
- text to match: core definition vs. examples included
- IR similarity measure in a word vector space: Dice coefficient, cosine, Jaccard, etc.
- word weighting function: none vs. *tf.idf*

Dispersal Match: For a (semantically related) cluster of concepts, consider the total set of candidate matches in the ontology. By picking one candidate for each concept, find the smallest (tightest, therefore most closely related semantically) cluster of candidates in the ontology. The variations include:

- full algorithm (combinatoric complexity) vs. greedy search
- number of candidates allowed per concept
- relative weighting of candidates from Name and Definition Matches

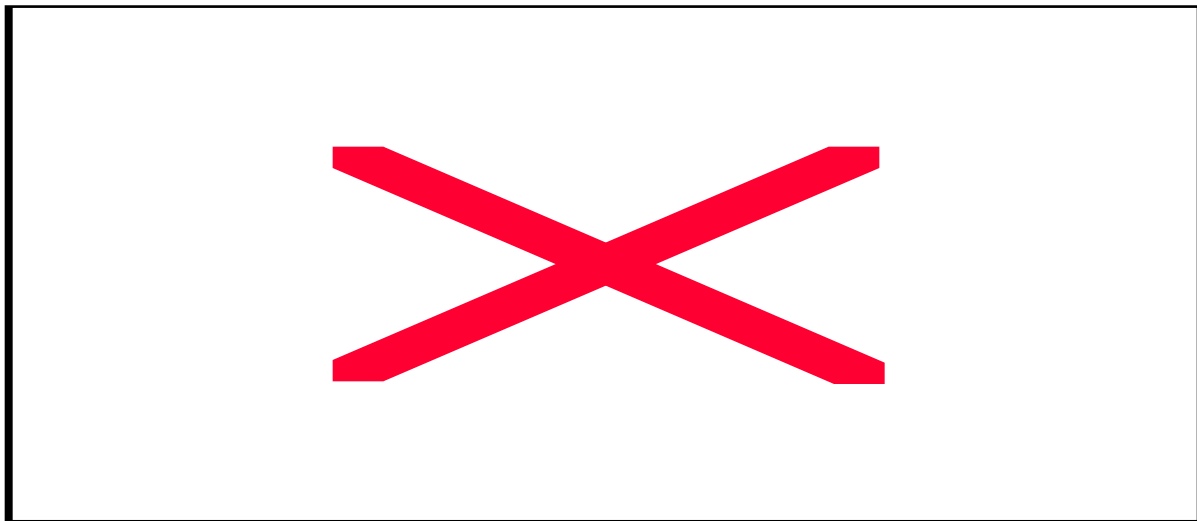
This match was invented for this project, and tested on a small set of 100 concepts, 10 candidates each, early in 2000. It performed very well [AKHP+00]. However, given the current number (7000) of concepts to be matched, we could not implement the full combinatoric algorithm. We therefore used a greedy search approach. Unfortunately, our experiments with this version of the Dispersal Match gave far less encouraging results. A more detailed description of our linking experiments can be found in [HPAA+01].

5. Data Warehousing for Efficient Evaluation of Decision Support Queries

A type of queries of particular interest are decision-support queries, also known as datacube queries, or on-line analytical processing (OLAP) queries. A datacube is a set of data measurements that are defined by the values of a set of dimensions in a way similar to our domain models. It can be thought as a multidimensional matrix. Datacube queries involve aggregations (sums, averages, etc, on groupings) of the data across one or several of the dimensions

In order to efficiently evaluate decision-support queries on our integrated statistical data, our architecture has a dual mode of operation. This extended architecture is shown in Figure 8. First, our system can retrieve live data from databases and web sources. This allows the users to obtain completely up-to-date data. However, for complex analytical queries that typically require large amounts of data and processing, live access does not offer the level of interactivity that some users require. Second, our system can warehouse the information from the data sources to allow for complex analytical queries to be executed much more efficiently. However, the data would be only as recent as the last update to the data warehouse. Figure 8(a) shows the process of loading the data warehouse. Once we have the sources modeled in our ontology we can retrieve all the available time-series and store them in a local data warehouse normalized under our common schema. On this local warehouse we can evaluate queries much more efficiently as we describe in the next section. Figure 8(b) shows how we propose to use the data warehouse. The user interacts with a friendly interface that passes formal queries (e.g. SQL) to the query planner (SIMS). Then the query planner analyzes the query and decides whether to retrieve the data live or to use the local data warehouse.

We are exploring several methods to efficiently evaluate decision-support queries once we the data is in our local data warehouses, see [ASSP01,RZ00] for details. In this paper we do not address the reasoning task that allows the mediator to decide whether to use the warehouse or the live sources.



(a) Loading the Data Warehouse

(b) Using the Data Warehouse

Fig. 8. Extending our integration system with a Data Warehouse

6. Conclusions

In this paper we have presented our integration framework, some initial results on semi-automatic domain modeling and source wrapping, and an approach to efficient evaluation of complex decision support queries with the aid of a local data warehouse. We have created a working prototype of the kind of system required to support information access over heterogeneous databases developed and maintained by different government and private-sector agencies.

Future work includes further development of our techniques for rapid inclusion of new databases into the system, enhanced term extraction and glossary mining, the development of sophisticated yet user-friendly interfaces tailored to the general public, and restricted but free-form natural language query input in multiple languages.

References

- [ACRR+94] 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.
- [AKH96] 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.
- [AK00] 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).
- [AAHP+01] Ambite, J. L., Y. Arens, E. Hovy, A. Philpot, L. Gravano, V. Hatzivassiloglou, and J.L. Klavans. Simplifying Data Access: The Energy Data Collection Project. *IEEE Computer* 34 (2), Special Issue on Digital Government, February 2001.
- [ASSP01] Ambite, J. L., C. Shahabi, R. R. Schmidt, and A. Philpot. Fast Approximate Evaluation of OLAP Queries for Integrated Statistical Data. *Proceedings of the First National Conference on Digital Government (dg.o 2001)*, Redondo Beach, May 2001.
- [BKMW89] 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.
- [F98] Fellbaum, C. 1998. (ed.) *WordNet: An On-Line Lexical Database and Some of its Applications*. Cambridge: MIT Press.
- [HRU96] Harinarayan, V., A. Rajaraman, and J. D. Ullman, 1996. Implementing Data Cubes Efficiently, *Proceedings of the 1996 ACM SIGMOD Conference*.
- [H98] 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.
- [HPAR00] Hovy, E.H., A. Philpot, J.-L. Ambite, and U. Ramachandran. 2000. Automating the Placement of Database Concepts into a Large Ontology. In preparation.
- [HPAA+01] Hovy, E.H., A. Philpot, J.-L. Ambite, Y. Arens, J.L. Klavans, W. Bourne, and D. Sarioz. 2001. Data Acquisition and Integration in the DGRC's Energy Data Collection Project. *Proceedings of the dg.o 2001 Conference*. Redondo Beach, California.
- [K88] Klavans, J. L. 1988. COMPLEX: A Computational Lexicon for Natural Language Processing. *Proceedings of Twelfth International Conference on Computational Linguistics (COLING)*. Budapest, Hungary.
- [KJT97] 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.
- [KM00] Klavans, J. L. and Muresan S. 2000. "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.

- [KW01] Klavans, J. L and B. Whitman 2001 “Extracting Taxonomic Relationships from On-Line Definitional Sources Using LEXING”
- [KL94] Knight, K. and S.K. Luk. 1994. Building a Large-Scale Knowledge Base for Machine Translation. *Proceedings of the AAAI Conference*.
- [M90] 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.
- [MMK98] Muslea, I. and S. Minton and C. A. Knoblock. 1998. Wrapper Induction for Semistructured Web-based Information Sources. Proceedings of the Conference on Automated Learning and Discovery. Pittsburgh, PA.
- [OH94] Okumura, A. and E.H. Hovy. 1994. Ontology Concept Association using a Bilingual Dictionary. *Proceedings of the 1st AMTA Conference*. Columbia, MD.
- [RA95] Rigau, G., and Agirre, E., 1995. Disambiguating Bilingual Nominal Entries against WordNet. *Proceedings of the 7th ESSLI Symposium*. Barcelona, Spain.
- [RZ98] Ross, K., and Zaman, K., 2000. Optimizing Selections over Datacubes, *Proceedings of the Statistical and Scientific Database Management Conference*.
- [SPKR96] Swartout, W.R., R. 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.

Acknowledgments

This work was funded by the National Science Foundation’s Digital Government Program under contract EIA-9876739.

Author Biography

Jose Luis Ambite is a senior research scientist at the University of Southern California’s Information Sciences Institute. His research interests include information integration, automated planning, databases, and knowledge representation. He received a Ph.D. in Computer Science from the University of Southern California.

Yigal Arens, co-principal investigator on the EDC project, is the director of the Intelligent Systems Division of the University of Southern California’s Information 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. He received a Ph.D. in Mathematics from the University of California, Berkeley.

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 semi-automated construction of large lexicons and terminology banks. He received a Ph.D. in Computer Science from Yale University. He is president of the ACL and past president of the AMTA.

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. She received a Ph.D. in Linguistics from the University of London.

Andrew Philpot is a scientific programmer at the University of Southern California’s Information Sciences Institute. He received an M.A. in Computer Science from Stanford University.