

# DGRC AskCal: A Multilingual Question Answering Agent for Heterogeneous Energy Databases

Eduard Hovy, Andrew Philpot, Lei Ding

DGRC

USC/Information Sciences Institute

4676 Admiralty Way

Marina del Rey, CA 90292-6695

{hovy,philpot,leiding}@isi.edu

Large collections of heterogeneous data, collected on the same domain by various government agencies at various times in various ways, can be very difficult for users—both experts and novices—to navigate through in order to find the data of interest. User difficulties result from several issues, including differences in terminology in the data collections, lack of familiarity with the terms in the domain, lack of expertise with the query mechanisms specific to particular end data sources, the need to combine or relate data from multiple locations in order to satisfy a single query, and the barriers posed when supplementary resources such as metadata, free-text documentation, and query construction tools are not accessible in a language the user understands well.

Tackling these problems, our research objectives are to:

- Enable novice users to perform sophisticated queries on heterogeneous government energy datasets using natural language (English, Spanish, and eventually Chinese);
- Develop general multimedia question-answering environment, allowing integration of language, ontology browsing, and menu filling, in order to form queries;
- Evaluate user preferences and effectiveness:
  - Task complexity from simple to complex,
  - User sophistication from novice to domain expert.

Since 1999 we have worked with members of the Energy Information Administration (EIA), Census Bureau, and Bureau of Labor Statistics (BLS), who have provided time series data about gasoline prices, in formats ranging from Oracle databases on CDs, text files, and live webpages.

Working with members of the DGRC at Columbia University, we have built Multilingual AskCal, which we will demonstrate at the conference. AskCal incorporates over 50,000 tables of data, homogenized under a 500-node domain model that is embedded in a 110,000-node ontology called Omega, built at ISI. The user can enter English or Spanish queries, whose lexical items are converted into domain terms via the ontology, and whose questions are converted into SQL queries for a query access planner, which can decompose them, draw data from various sources, and recombine them.

This research has achieved goals in various directions:

- Built the AskCal prototype, working in English and Spanish, accessing a large amount of data, demonstrated over several years;
- Built a data conversion tool using some of this technology in conjunction with Fetch Inc., and delivered this to the EIA as a paid commercial product;

- Built several supporting tools and prototypes, in areas as diverse as text processing, data access, multimedia interfaces, and user studies (some of this at Columbia);
- Supported theoretical research in several areas, including advances in rapid database access (Ken Ross and students at Columbia); automated ontology/model alignment (Hovy and colleagues at ISI); glossary harvesting and domain modeling: (Klavans and students at Columbia, and Hovy and colleagues at ISI); multiple-interface access to databases (Feiner and students at Columbia and Hovy and colleagues at ISI), resulting in approx. 20 journal or conference papers and numerous presentations;
- Established an effective working collaboration between the two partners of DGRC, namely USC/ISI and Columbia University.

During the course of this work, we encountered several barriers, some of them quite serious. For fully understandable reasons, especially the government's privileged position when it comes to citizen privacy, it is difficult for government partners to provide raw data to researchers. However, raw data is precisely what research projects need in order to be most helpful, because that's where most of the problems lie! A second important issue is the lack of discretionary funds in government partners, which limits the size and extent of transferred technology. This compounds the difficulty in defining research problems that interest both researchers and the government partner; the researchers are too easily seen as handy free software builders. In fact, a general difference in culture, language, and expectations between researchers and government employees makes especially the startup period difficult, which is a pity, since we believe it central for successful collaboration that good and frequent contact take place.

Nonetheless, we feel privileged to have been able to work on this research. The government offers interesting, real-world problems—the research challenges are not artificial and point to problems that pure theoreticians might overlook (e.g., 'traditional' query interfaces to databases are of limited practical utility)—and a chance to develop technology that will make life better for the citizen, not just exist on paper.

For the future of the NSF's Digital Government program, we offer two recommendations:

- Always start with small 9-month pilot projects; they:
  - establish deeper research–government connections,
  - allow researchers to really understand the government's needs,
  - allow researchers to build small pilot testbeds,
  - enable government to get permission to provide data.
- Privately educate government partners not to expect working systems, but to look for longer-term benefits.