

2001 Project Report to NSF

Energy Data Collection Project

Year 2

**Jose-Luis Ambite* and Yigal Arens* and Walter Bourne† and Steven Feiner† and Eduard Hovy*
and Judith Klavans† and Andrew Philpot* and Ken Ross†**

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

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

Contacts: Eduard Hovy, ISI
hovy@isi.edu

Judith Klavans, Columbia
klavans@cs.columbia.edu

Abstract

The massive amount of statistical and text data available from Federal Agencies has created a set of daunting challenges to both research and analysis communities. These problems include heterogeneity, size, distribution, and control of terminology. At the Digital Government Research Center we are investigating solutions to three key problems, namely, (1) ontological mappings for terminology standardization; (2) data integration across data bases with high speed query processing; and (3) interfaces for query input and presentation of results. This collaboration between researchers from Columbia University and the Information Sciences Institute of the University of Southern California employs technology developed at both locations, in particular the SENSUS ontology, the SIMS multi-database access planner, the LKB automated dictionary and terminology analysis system, and others. The pilot application targets gasoline data from BLS, EIA, Census, and other agencies.

1. EDC Project Overview

The massive amount of statistical and text data available from Federal Agencies has created a set of daunting challenges to both research and analysis communities. These problems include heterogeneity, size, distribution, and control of terminology. At the Digital Government Research Center we are investigating solutions to three key problems, namely, (1) ontological mappings for terminology standardization; (2) data integration across data bases with high speed query processing; and (3) interfaces for query input and presentation of results. This collaboration between researchers from Columbia University and the Information Sciences Institute of the University of Southern California employs technology developed at both locations, in particular the SENSUS ontology, the SIMS multi-database access planner, the LKB automated dictionary and terminology analysis system, and others. The pilot application targets gasoline data from BLS, EIA, Census, and other agencies.

The DGRC Energy Data Collection (EDC) Project was started in the National Science Foundation's Digital Government program in 1999. It is developing solutions to three key problems in accessing large distributed data collections: (1) sophisticated planning of access to multiple distributed and heterogeneous databases; (2) the use of a large ontology for terminology standardization and user guidance; (3) flexible user interfaces for query input, ontology browsing, and result presentation. A short review paper is included at the end (Ambite et al., 2001).

In this work, the proposing team is working with representatives of major Federal and State statistics agencies and other organizations and individuals on a regular basis, to collect and disseminate statistical data. Representatives of these agencies, primarily from the Census Bureau, the Bureau of Labor Statistics (BLS), and the Energy Information Administration (EIA) of the Department of Energy (DoE), and the California Energy Commission (CEC). Other agencies we have met with include the National Center for Health Statistics (NCHS) and the Los Angeles County Administration. The Energy Information Administration provides extensive *monthly energy data* to the public on its Internet site (<http://www.eia.doe.gov>). The site is heavily browsed, receiving hundreds of thousands of hits a month, even though most of the information is available only as downloads of standard web (HTML) pages or as prepared PDF documents. Monthly data can only be obtained for the last few years in this manner. Current facilities thus provide only limited access to this very rich data source. Some portion of the data is also accessible by querying, but there are two serious problems hampering the current query system. First, it does not provide visibility for the many definitions and footnotes that explain the complex nature of the data and to changes that occur in series over time. Lack of awareness of such explanatory information often makes incomparable figures appear to be comparable. Second, the difficulty of defining queries makes the querying system useful only to expert users.

The EDC Project is addressing both problems. Techniques are being developed and implemented to attempt to make the complexities of data series either transparent, or more visible to users, depending on whether they can be handled independently by the system or not. And novel query facilities and other data analysis and presentation capabilities are being developed, that will be usable by the more common potential user of EIA data—the non-expert. Our aim is to build a system that, though still a research prototype, will be of benefit to various segments of the public, judging by the high demand for monthly energy data. Besides the large number of browsers accessing EIA's Web site, hardcopy publications further disseminate EIA's monthly energy data. Such paper publications include the Monthly Energy Review and the International Petroleum Statistics Report. We will make the monthly data available through EDC's new query system on the EIA's web site.

In the first year of R&D, we have focused on three areas:

Information Integration. The research questions we have addressed include effective methods to identify and describe the contents of databases so that useful information can be accurately and efficiently located even when precise answers are unavailable. We have wrapped over 120 databases for testing the first stage of information integration, performed research on computational properties of aggregation, and investigated the extraction of information from footnotes embedded in text.

Ontology Construction. We have built on the large formal SENSUS ontology at USC/ISI currently containing over 90,000 terms, linked together into a subsumption (*isa*) network, with additional links for *part-of*, *pertains-to*, and so on. Our results include extending the SENSUS high-level general domain ontology to incorporate new domain models and extending and developing new automated concept-to-ontology alignment algorithms. Term extraction from glossaries involves the automatic analysis of 6000 terms across agencies (EIA, Census SICS and NAICS codes, EPA) and the automatic handling of acronyms towards the creation of a cross-agency ontology.

User Interface Development. We have designed and implemented a user interface with the capability of handling integrated querying and presentation of results.

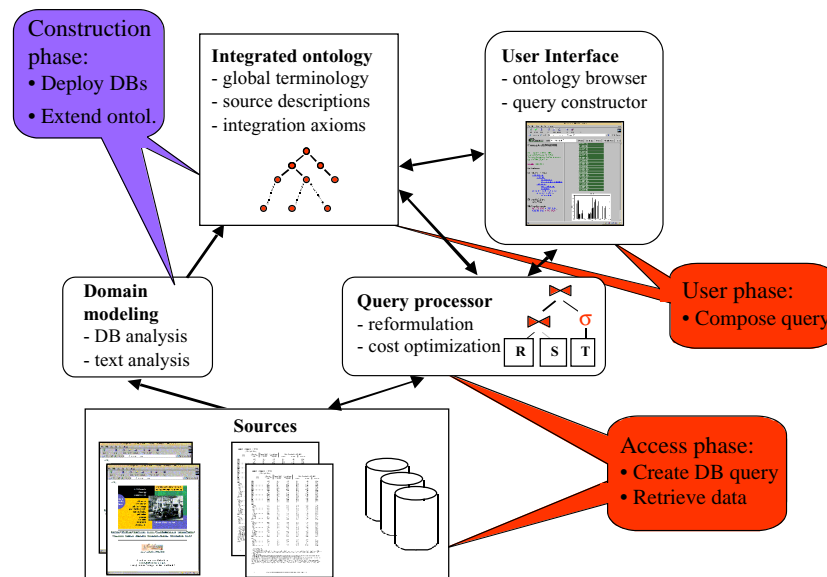


Figure 1. Architecture of the EDC System.

In order to support homogeneous access to multiple databases, the EDC system (architecture in Fig. 1) involves three principal components: the database manager and access planner, the overarching ontology in which terms are defined and standardized, and the interface. Work relating to the database manager and access planner involves locating, preparing (“wrapping”), and incorporating new databases into the system. Central to this enterprise is the construction of models of the data, which are used for access planning. Ontology-related work focuses on the construction of models of the domain and the embedding of these models into the ontology. This includes work on semi-automated term alignment, term extraction from glossaries, and acronym handling. The interface supports access to this data by allowing the user to formulate queries in a variety of ways: either by filling in menus, or by browsing the domain and data models via the ontology, or, eventually by entering a question in restricted English.

2. Recent Work

Workplan for last half of 2000

This is the plan of work we developed for the last half of 2000.

1. Data integration

1.1 Current data model, and databases incorporated to date.

1.2 Recent delivery of databases to EIA

1.3 Semi-automated data and domain modeling

2. Ontology extension

2.1 Ontology renewal

2.2 Clustering and alignment research, and glossary linkage (with Columbia)

2.3 Learning instantial knowledge from text

3. Lexical extraction of information from on-line glossaries

Extraction of information for domain modeling and ontology construction

4. System architecture work

4.1 SIMS issues

4.2 Port to Linux on new PC

4.3 Port to new Lisp

4.4 Smooth down rough edges remaining from demo in May

5. User testing

Design and apply tests on usage of the system

6. User interface

Design of new user interface

7. Database query and results management

Very fast access to large sets of precompiled data

The remainder of this section provides more detail.

2.1 Data Integration (ISI: Philpot)

Current Data and Data Models

The following data has been captured in the system to date.

A. BLS

There are 53 BLS (Bureau of Labor Statistics) series, all of which are obtained by a wrapper which instantiates a form on the BLS web site.

- 27 of these are BLS/WP "PPI Revision Commodities"
- 12 of these are BLS/PP "PPI Revision Current Series"
- 8 of these are BLS/AP "CPI Average Price Data"
- 6 of these are BLS/CU "CPI All Urban Consumers"

Each BLS series provides:

BLS-PERIOD	string	encoded field for quarterly, monthly
FOUR-DIGIT-YEAR	string	year of measurement
ORIGINAL-SERIES-NAME	string	semantic name of series if avail
SERIES-ID	string	encoded name of series if avail
SOURCE-TAG	string	obsolete
SOURCE-URL	string	URL where source obtained
STANDARD-DATE-NUMBER	string	date from year+period in YYYYMMDD fmt
STANDARD-DATE-STRING	string	date of measurement
SVALUE	string	value of measurement, as string
VALUE	number	value of measurement

B. CEC

There are 3 CEC (California Energy Commission) series, which are obtained by wrapping a single page on the CEC web site.

CONVENTIONAL-USA-DATE	string	date stored like 1/2/99
ORIGINAL-SERIES-NAME	as above	
SERIES-ID	as above	
SOURCE-TAG	as above	
SOURCE-URL	as above	
STANDARD-DATE-NUMBER	as above	
STANDARD-DATE-STRING	as above	
VALUE	as above	

C. EIA

There are 16 EIA (US DOE Energy Information Administration) 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-PERIOD	string	period (cleaned up via program)
FOUR-DIGIT-YEAR	as above	
ORIGINAL-SERIES-NAME	as above	
RAW-PERIOD	string	period (freq) as extracted
SERIES-ID	as above	
SOURCE-TAG	as above	
SOURCE-URL	as above	
STANDARD-DATE-NUMBER	string	computed from year+EIA period
STANDARD-DATE-STRING	as above	
SVALUE	as above	
VALUE	as above	

D. EIA OGIRS

There are 28 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; as well, we have imported the data themselves into Oracle. OGIRS itself has thousands more series which we anticipate including in a later phase. OGIRS also has much more formal metadata that can be viewed in Access, but retrieving it is not straightforward.

OGIRS-DATE	string	date in OGIRS-specific format
ORIGINAL-SERIES-NAME	as above	
SERIES-ID	as above	
SOURCE-TAG	as above	
SOURCE-URL	as above	
STANDARD-DATE-NUMBER	string	date computed directly from OGIRS format
STANDARD-DATE-STRING	as above	
SVALUE	as above	
VALUE	as above	
VB-DATE	number	date using Visual Basic epoch

Besides the time series ("MEASUREMENT") concepts detailed above, the EDC demo also provides footnote information, by joining through parallel ANNOTATION and FOOTNOTE concepts, which share the definitional metadata above but have a few different retrievable attributes. They are omitted here.

All this data is organized into a single Data Model, which in turn is embedded into SENSUS. The Data Model has the following primary dimensions (data 'columns').

The primary key for MEASUREMENT is {SERIES-ID STANDARD-DATE-NUMBER} which means that a SERIES-ID must be found or imposed for each source that encodes all identifying geographic, frequency, product, etc., information. Generally such IDs are available naturally in the source. Additionally, as an engineering practice, we have imposed 10 definitional attributes onto every domain concepts. Each

measurement concept thus possesses all of the following attributes, which can be retrieved as if they exist in an end source.

AGENCY-NAME	string	agency providing data
AREA-NAME	string	locale of data
FAMILY-NAME	string	subgroup of series within agency
FREQUENCY-NAME	string	how often/when measured, e.g. monthly
POINT-OF-SALE-NAME	string	where in the supply chain measured
PRODUCT-NAME	string	what product e.g. unleaded regular
PROVENANCE-NAME	string	how data obtained: web, RDBMS, etc.
QUALITY-NAME	string	what kind of measurement: vol, price
SEASONAL-ADJUSTMENT-NAME	string	is data seasonally adjusted
UNIT-NAME	string	units of measurement: Mbbl/mo, etc.

We have recently added 29000 new EIA OGIRS series to SIMS.

Recent delivery of databases to EIA (ISI: Philpot, Ambite)

After discussions with Cal Kilgore of EIA it became clear that some of the data EIA would like to include in their monthly publication is not in an appropriate format (it mixes states' petroleum data per month).

In particular, the EIA would like to refer from indexes of single-state data sets (e.g., http://www.eia.doe.gov/emeu/states/main_ca.html) to relevant subsections of various text documents containing composite information, (e.g., just the California portion(s) of http://www.eia.doe.gov/pub/oil_gas/petroleum/data_publications/petroleum_marketing_monthly/current/txt/tables31.txt).

Since data changes only weekly, a natural approach here was to generate snapshot breakout reports each weekend of each desired composite report, and publish them to the web.

Using ISI's existing expertise and technology we were able to download, reformat, and deliver to EIA the data in a form they were able to deal with. Handling over 60% of their data type took less than a week.

In detail, this process required three steps:

1. Wrapping the data. By hand, we classified each text report file's structure into one of three forms. We then used ISI's Ariadne table interpretation technology to extract a series of pages from each multi-page report according to the (form-specific) characteristic set of text landmarks and text entries. First we identified left- and right- hand pages when present, then we broke each page into header, body, and footer (with footnotes). Headers are wrapped to extract type/subtype and measurement/column labels. Bodies are wrapped to extract information in columns, using landmarks, normal text conventions, and apparent column widths. Table conventions, landmarks, and limited amount of domain-specific information is used to repair multi-line tokens and to associate subtotals with their superheadings. The resultant relation is saved in a neutral interchange format.

(For the example table 31 above, the result is

<http://www.isi.edu/~philpot/eia/form1/monthly/tables31.txt.csv>

This representation "flattens" the original tables into 7-tuples:

(product,frequency,locale,date,subtype,measurement,value).

2. Posing appropriate queries against the wrapped data. Henry Weigel from the EIA gave requirements for per-state breakouts from composite reports, which effectively require database selections, projections, and potentially unions between the relations. In a completed system, data in the interchange format from step 1 above would be imported into a RDBMS or integrated by an information source mediator such as SIMS. At this point, we implemented a simple emulation that supports only conjunctions of selections and projections. Using this command-line interface, we generated per-state breakouts of each measured product. The results are stored as simple HTML tables with little to no formatting, intended as an intermediate visualization tool rather than an end-user accessible report. An example of this data is

http://www.isi.edu/~philpot/eia/form1/monthly/monthly_tables31.txt_California.html

3. HTML presentation of the results. This is a more open-ended process than the above. Our preliminary discussions with Henry Weigel from EIA suggest that it will be sufficient to re-integrate with an HTML rendition of the original headings, effectively to splice the original heading together with only the rows relevant to a state-based breakout. We will probably want to honor the original heading/subhead as much as possible, although subhead totals may no longer be relevant. Finally, we hope to parameterize the output using HTML Cascading Style Sheets instead of significant embedded style tags, to reduce file size and support future malleability of styles.

Initial results can be seen at

http://zeus.isi.edu/eia/tables31.txt_District_of_Columbia_monthly.html.

We are including this new data in our system's collection of databases as well.

Semi-automated data and domain modeling (ISI: Hovy, Ambite, Philpot)

We have done some preliminary thinking about creating code to help automate the process of creating Data Models. This work involves using 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. Since this is very ambitious, we broke down the work into a series of steps.

Given some new domain:

1. Text: create concepts from the textual data surrounding the data
2. Databases: create concepts from the databases and metadata themselves
3. Ontology: (cluster and) ontologize these concepts into SENSUS
4. Domain Model: (manually) use this information to build a domain model

The following provides more detail.

1. Text Analysis/Information Extraction

From glossary definition to a structured template. Templates include standardized items such as dates, dollar amounts, etc., which Columbia's methods can extract, and more complex NPs and so on, which involve linking to the ontology, which we do.

2. Database Concept Extraction

We have listed some ideas and a procedure, and found that similar work is being done in the data mining community and at U of W Seattle. Input sources include:

- data (numeric; string)
- metadata (table and column names; text (including footnotes, glossaries, and text))
- Ontologies (terminologies (Sensus); logical (CYC))
- partial Domain Models

Techniques include:

- maximal separation: every attribute in sources is a concept
- structuring (called aggregation in OO literature):
- attributes of a source table are related (may form a class, or may belong to different classes along a hierarchy)
- value/value comparison (compare values in two attributes of two source tables (contained, overlap, disjoint; constraint can be learned on those; most common case)
attribute-name/value comparison (sometimes the values of an attribute dictate a partition of different tables in another source)
- table-name/value comparison (and same for table names)

Possible characterizations of a single attribute include ranges, enumerated sets, minimum, maximum, average values, orthographic patterns ex: (310) 740-1223, typing (numbers, letters), and information theory.

Atoms include cell values, attribute names, table names, and source names.

3. Ontology Matching

We are developing a sequence of steps:

a. For a given (set of) concepts, propose candidate equivalents or near-equivalents in SENSUS. Match algorithms:

- Name Match
- Definition Match
- Structure Match

exploit relationships among concepts (ex.: if things have a price then they must be of the classes of things that have prices)

b. Given a set of candidate matches, rank them and select the best one(s)

- Dispersal Match (validation): cluster closely related concepts, then prefer some candidates based on cluster tightness
- other? Automated learning of relative strengths of matchers. Combination of matchers using stacking, Muslea's methods, etc.

More details appear below.

4. Domain Modeling.

Perhaps this involves an interface like the one developed in EXPECT.

5. Testing

To validate/evaluate test these ideas, we will use a new database.

2.2 Ontology extension (ISI: Philpot, Ambite)

Ontology renewal: New version of SENSUS using WordNet 1.6 and new Lisp

We created a new version of SENSUS, out of WordNet 1.6 (from Princeton). This work involved a fresh generation of concept names using a modified version of the Graehl algorithm, and some reorganization of some new concepts under the Upper Model.

We have deployed this version of SENSUS, complete with EDC Domain Model, on a new platform, on a PC (under Linux), as well as under Sun Solaris. The new SENSUS (and its representation language SENSOR) has been ported to ACL 5.0.1.

We have not yet deployed the new SENSUS under Ontosaurus (the web browser for SENSUS). Given new developments in Lisp (the language of SENSOR, the implementation of SENSUS), it makes sense for us to reimplement Ontosaurus, using as much of the old code as possible.

We are planning to port the new SENSUS to Columbia.

We have also created the following API for concept definitions.

SENSUS CONCEPTS

These are the major slots for concept definitions (although any new slot can be added):

:DEFINITION	<string>
:DIRECT-SUPERCLASS	(<concept>+)
:PART-OF	(<concept>+)
:MEMBER-OF	(<concept>+)
:SUBSTANCE-OF	(<concept>+)
:PERTAINS-TO	(<concept>+)
:WN-TYPE	<wn-type>
:HAS-PARTS	(<concept>+)
:HAS-MEMBERS	(<concept>+)
:HAS-SUBSTANCE	(<concept>+)

where <wn-type> is one of the SENSUS types, e.g., NOUN.ANIMAL, VERB. PERCEPTION, etc. (a lisp symbol), and inverses of the above may also be used if needed.

The following are new attributes:

:EXAMPLES (<string>+)
 :WN-OFFSET integer
 :SOURCE string or symbol

Example:

```
(DEFCONCEPT |bank,side|
:DEFINITION "an elevated geological formation"
:EXAMPLES ("he climbed the steep slope"
           "the house was built on the side of the mountain")
:DIRECT-SUBCLASS (|hillside|)
:DIRECT-SUBCLASS (|mountainside|)
:DIRECT-SUBCLASS (|canyonside|)
:DIRECT-SUBCLASS (|downslope|)
:DIRECT-SUBCLASS (|acclivity|)
:DIRECT-SUBCLASS (|bank=sloping land (especially|)
:DIRECT-SUBCLASS (|cant,bank|)
:DIRECT-SUPERCLASS (|geology,formation|)
:PART-OF (|natural elevation|)
:SOURCE :WORDNET1.6
:WN-OFFSET 6724958
:WN-TYPE NOUN.OBJECT
:?WN-SENSE-KEYS ("incline% 1:17:00::" "slope% 1:17:00::" "side% 1:17:01::")
:?WN-POS :NOUN)
```

SENSUS WORDS

These are the major slots for words:

```
:SENSES ((<concept> <pos> <sense-number>)+)
:SOURCE as above
```

where POS is one of NOUN, VERB, ADJECTIVE, ADVERB where SENSE-NUMBER numbers the senses for each POS in order of commonness.

Example:

```
(DEFWORD S2.T::|incline:WORD|
:SENSES ((|bank,side| :NOUN 1)
         (|incline,ramp| :NOUN 2)
         (|tend<be| :VERB 1)
         (|tend<think| :VERB 2)
         (|slope<tip| :VERB 3)
         (|make willing| :VERB 4))
:SOURCE :WORDNET1.6)
```

Clustering and alignment research, and glossary linkage (ISI: Hovy, Philpot, Ambite)

We spent a significant amount of time investigating how to automate the linking of concepts into the ontology.

Our initial work on this was reported in the project report at the May 2000 dg.o

workshop. At that time, we had only about 100 concepts to link into SENSUS.

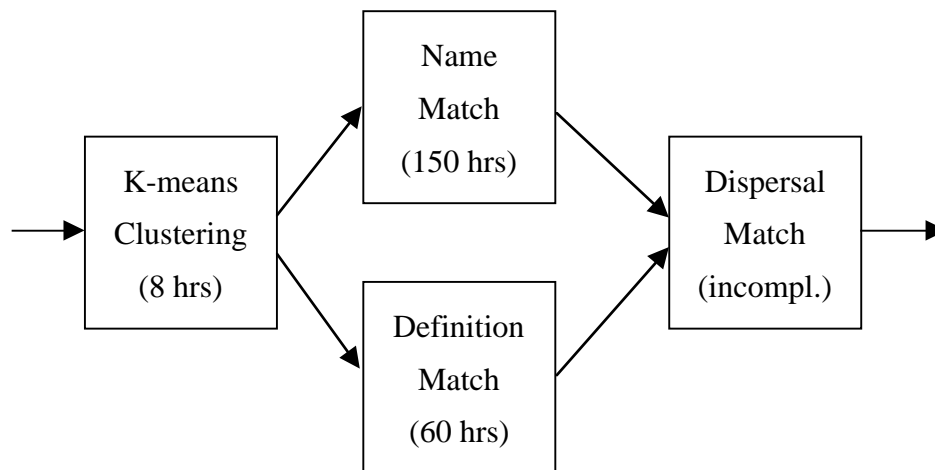
Since then, Columbia University has extracted approx. 6,000 concepts from glossaries and other text acquired from EIA and BLS. Our challenge was to link these concepts into SENSUS at their most appropriate points.

We extended and made more robust the methods we had developed with a graduate student in the first half of 2000. We implemented new concept clustering algorithms. We significantly improved the speed of string matching algorithms. This work is almost complete, but still needs evaluation of the results.

The concept linking process has the following steps:

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 following matching algorithms to each concept, and collect for each concept the candidate matches in SENSUS:
 - Name Match (various variations)
 - Definition Match (various variations)
4. apply the Dispersal Match algorithm 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

Times given are for handling the 6000 glossary concepts from Columbia University.



We developed several variations of the various algorithms, and investigated the performance of each.

Clustering:

Definition: for a given set of concepts, cluster them to find the sets of ones that belong together semantically. Traditionally, 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, on 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:

Definition: match the name of the concept against the name of each ontology concept and rank 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:

Definition: 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 vector space: Dice coefficient, cosine, Jaccard, etc.
- word weighting function: none vs. *tf.idf*

Structure match: (not implemented)

Definition: for the concept and each ontology concept, match their related concepts (for a given relation) using name and definition match. The relations include:

- superclass/subclass (hypernymy/hyponymy)
- part-of
- etc.

Dispersal Match:

Definition: 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. However, given the current number (6000) 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. We therefore decided to test the performance of the whole sequence on ideal data, in order to determine where the problems arose.

To create the ideal data, we extracted from the ontology four subhierarchies (furniture, beverages, tools, and motor vehicles). We applied the match sequence to these hierarchies in order to see how well they get matched to their true locations. We expected that the Dispersal Match would find the best matches easily, but this turned out to be not the case. The confusion matrix in table 1 shows that the results, though much better than random, are in many cases still around 65%.

One reason was inadequate matching, which produced spurious candidate matches. For this reason, we decided to upgrade SENSUS to WordNet 1.6, which includes both a more logical internal organization and better and more complete concept definitions. At the time of writing we have completed the transition to WordNet 1.6.

The other reason is inaccurate clustering. We will continue with the matching work later this semester.

Tools	Beverages	Furniture	totals
Global % distribution			
5.90%	5.76%	6.91%	18.56%

35.83%	4.03%	17.84%	57.70%
0.72%	22.73%	0.29%	23.74%
42.45%	35.52%	25.04%	100.00%
Row distribution			
13.90%	17.70%	27.59%	
84.41%	12.39%	71.26%	
1.69%	69.91%	1.15%	
100.00%	100.00%	100.00%	
Column distribution			
31.78%	31.01%	37.21%	100.00%
62.09%	6.98%	30.92%	100.00%
3.03%	95.76%	1.21%	100.00%

Table 1: Confusion matrix showing results of concept matching on three sets of ideal data. Ideally, one cell in each row (or column) will contain 100%, and the others in that row (or column) 0%.

Learning instantial knowledge from text (ISI: Michael Fleischman (student) and Hovy)

We have been experimenting with techniques to extract instantial knowledge (specific instances of things in the world, not general concepts) from free text. The idea is to extend SENSUS to know things about the world, starting with simple things such as names of cities, countries, movie stars, etc.

There has been much interest in the recent past concerning the automated categorization of elements within text, such as named entities. We focus on developing methods for the subcategorization of location names. Subcategorization of locations is not a trivial task even for human subjects, who perform at accuracy levels of less than 58%. However, after experimenting with both Bayesian classifiers and decision tree learning algorithms, we have designed a system that achieves accuracy levels greater than 80%.

The challenge is to decide whether the following items are *cities*, *regions*, *mountains*, *rivers*, *states*, or *territories*:

- 1 " This destructive virus is spreading world-wide very rapidly , " read one January posting on a *Vsaogptmos*_____ board.
- 2 "Now people are fantasizing about *Dpiyj Zommrdpys* _____ going up in ashes because the plutonium is somehow wired into the computer , " he says.
- 3 The pulp and paper operations were moved to *Dpiyj Vstpaoms* _____ in 1981 .
- 4 The company , which is based in *Dpiyj Dsm Gtsmdodvp* _____, *Vsaog.* _____ said an antibody prevented development of paralysis in about 70% of the treated rats , and delayed and reduced the degree of paralysis in the other cases .
- 5 The *Fstorm* _____ headquarters employs fewer than 70 people .

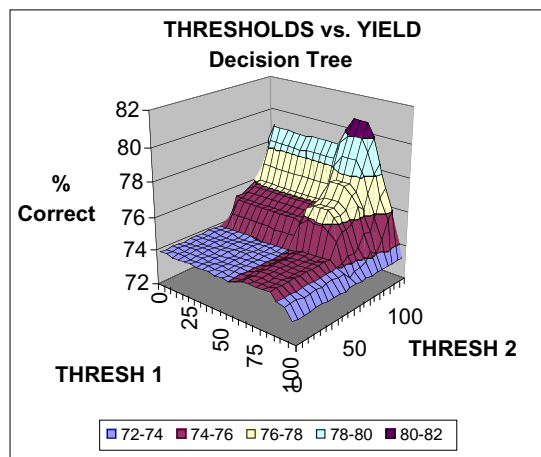
	/ # Total	Correct	/ # Total	Correct
City	260 / 373	69.7	309 / 373	82.8
Country	411 / 482	85.3	440 / 482	91.2
Street	6 / 10	60.0	6 / 10	60.0
Region	44 / 65	67.7	44 / 65	67.7
Water	6 / 7	85.7	6 / 7	85.7
Artifacts	4 / 8	50.0	4 / 8	50.0
Territory	71 / 148	48.0	79 / 148	53.4
Mount	0 / 3	00.0	0 / 3	00.0
Total	802 / 1096	73.2	888 / 1096	81.0

The Decision Tree, augmented with MemRun, outperforms human judges who were tested on a subset of the same task.

Human judges

	Subject 1		Subject 2		Subject 3		Avg.
	# Correct	%	# Correct	%	# Correct	%	%
Class	/ # Total	Correct	/ # Total	%Correct	/ # Total	Correct	Correct
City	5 / 7	71.4	5 / 7	71.4	4 / 7	57.1	66.7
Country	27 / 37	73.0	25 / 37	67.6	17 / 37	45.9	62.2
Street	2 / 5	40.0	2 / 5	40.0	1 / 5	20.0	33.3
Region	4 / 8	50.0	4 / 8	50.0	3 / 8	37.5	45.8
Water	0 / 3	0.0	0 / 3	00.0	0 / 3	00.0	00.0
Artifacts	2 / 2	100.0	0 / 2	00.0	2 / 2	100.0	66.7
Territory	8 / 11	72.7	5 / 11	45.5	10 / 11	90.9	69.7
Mount	1 / 2	50.0	0 / 2	00.0	0 / 2	00.0	16.7
Total	49 / 75	65.3	40 / 75	53.3	40 / 75	53.3	57.3

Varying the values of THRESH1 and THRESH2 gives the following result:



This graph shows an interesting relationship between the confidence values of the rules entered into the temporary database and the confidence values of rules that should be checked by the database. The graph shows that the optimal yield is not produced by entering and checking *all* classifications as was done in experiment 2. Rather, it is the configuration mentioned above which produces the highest overall accuracy. This is not surprising, as such a configuration would avoid the contamination of the database by poor classifications, as well as not risk a misclassification when the classifier is very certain of its initial hypothesis. This data implies that better results can be obtained with the Bayesian classifier as well, and will be examined in future research.

The work is being done by Eduard Hovy and a graduate student, Michael Fleischman. A paper was accepted at the ACL Student Session.

3. Lexical extraction of information from on-line glossaries (Columbia: Klavans and students)

In order to achieve the goal of extracting and structuring conceptual knowledge from glossaries into the SENSUS ontology, we have identified three distinct processes:

- a. identification of definitional material from web pages or other online sources
- b. parsing of definitions for salient properties and features
- c. incorporation of the structured information into a larger ontology, including linking and merging of definitions from different agencies and sources

In year one, we built a system called LEXING to extract LEXical INformation from Glossaries. LEXING took as input a set of dictionary files and output a set of structured concepts. Thus, our primary focus in year one was on the middle or parsing step of the three step process outlined above.

In year two, we made progress on all three aspects of the problem. First, we improved our webcrawler to function over .gov sites to identify and extract glossaries and definition files. Many of these files are in a variety of formats, not only .html and .pdf but also one definition to an .html page or many to a page, and so on. The routines to clean and normalize these files will be part of our work in year three.

Second, we improved the parser so that we now have a series of routines to correctly handle empty and common head or genus terms within a particular domain, and to link prepositions for use in skipping past empty head genus phrases. We have completed an evaluation of the definition content analysis which shows that, over and EIA, EPA and Medical data set we find from 92% - 98% genus phrases, a range from 5%-79% properties, and up to 39% of the quantified phrases. As for accuracy, we ran LEXING over five different sets of definitions looking only for accuracy of genus term identification. We manually tagged 500 definitions, 100 from each of the five domains of Civil Engineering, Computer Terms, Biomedical Information, General Medical Information and Energy Information. To compute scoring for genus term identification, a match was defined as both the manual tagger and LEXING choosing the same genus. The accuracy results were:

Civil Engineering	94%
Computer Terms	80%
Biomedical Information	92%
General Medical Information	97%
Energy Information	83%

An analysis of the errors revealed that future research needs to include some statistical analysis of frequently occurring heads to improve the head-preposition linking modules. Furthermore, although most definition files follow patterns typical of glossary entries, some entries had non-standard verbal and sentential elements that require careful pre-processing.

2.4 System architecture work (ISI: Philpot and Ambite)

SIMS issues

We have investigated and improved the performance of SIMS in a variety of ways, including adding the capability to cache queries, and, using Allegro ODBC, to remove the http/ODBC broker agent.

Port to Linux on new PC

We are porting the entire system to run on a PC platform, under Linux. This involved extensive reengineering of certain parts of the code. In particular, SIMS was ported to run on a different kind of Lisp.

Port to new Lisp

We have ported the system to ACL 5.0.1.

2.5 User Interface (Columbia: Steven Feiner and student)

Our focus in the second year of the project has been on the improvements and changes to the initial user interface in preparation for robust, portable, and efficient user interface that facilitates user access to data from multiple sources/agencies. The initial interface addressed the following main tasks: support of adaptive, context-sensitive queries via a system of guided menus; display of tables created by the integration back-end from one or multiple individual databases, along with footnotes and links to original data sources; and browsing of the ontology that supports the entire integration model, with the capability to display concept attributes, relationships, and definitions in graphics and text.

The new approach that we have developed focuses on developing the query interface in a way that links it with the ontology. We have used principles of cognitive science and visual focus to build a new way to view complex data. Instead of the “peep-hole” confusions which are typical of walk-through menus, our new “spotlight” view permits

the user to highlight details without losing their current location or obscuring other related data. We eliminate the kind of pop up menu that often clutters displays, thus keeping a clean but thorough view of the data. We have also set up a machine with a local version of the full SENSUS ontology, so that we can more rapidly test user reactions.

2.6 Database query and results management (Columbia: Ken Ross)

We have adapted the use of the datacube for FedStats data to compute aggregates of a set of database records at a variety of different granularities. An example of a datacube on census data might be: Broken down by age, rating of English proficiency, and number of children, find the maximum income including subtotals across each dimension. This implies that in addition to computing the maximum income over all the data, we would have also have to compute the maximum income over all ages, all English proficiency ratings, all (age, proficiency) pairs etc. Since we have 3 dimensions there are 2^3 granularities at which we would have to compute subtotals.

Datacubes are useful for many kinds of data analysis in which one can “slice and dice” multidimensional data to observe patterns. Such an approach is useful in many digital government contexts, including any large multidimensional data set over which aggregates are needed at various granularities. Datacubes can also be used to analyze data for privacy issues. For example, the Census bureau has the responsibility for disseminating detailed data about the U.S. population without revealing data so detailed that individuals can be identified. An analysis of the datacube can help identify cells in which the population is too small for safe disclosure of the detailed data; only aggregated data would be released in such a case. In order to support on-line access to datacube results, one would like to perform some precomputation to enhance query performance.

We provide a main memory based framework that provides rapid response to queries and requires considerably less maintenance cost than a disk based scheme in an append-only environment. For a given datacube query, we first look among a set of previously materialized tuples for a direct answer. If not found, we use a hash based scheme reminiscent of partial match retrieval to rapidly compute the answer to the query from the finest-level data stored in a special in-memory data structure.

We use a two-level materialization scheme. Our level-1 store contains datacube tuples. Since we cannot expect to materialize all datacube tuples, we store only “high value” tuples in the level-1 store. We analyze what constitutes a “high value” tuple, and demonstrate that tuples at coarse granularities and tuples with high query probabilities are good candidates for materialization.

Our level-2 store contains all tuples at the finest granularity. We store the data in a structure that allows a form of partial match query to answer queries without scanning the entire finest granularity dataset. We interrogate the level-2 store only if we find no match in the level-1 store. Our data structures enable fast incremental updates in response to new data, thus allowing the datacube server to supply up-to-date results.

Our approach yields rapid query responses for the important class of applications in which the finest granularity tuples of the datacube fit in main memory. Rapidly decreasing main memory prices have lead to workstations with over a gigabyte of RAM being commonplace today.

Finally, our techniques can be used by analysts inside government agencies to test whether their data conform to privacy requirements. This is particularly important for agencies such as Census which must be careful not to reveal personal information.

3. Plans for the Future

3.1 Workplan for First Half of 2001

This is the plan of work for the first half of 2001.

Ontology

Terminology Extraction (Columbia: Klavans and Sarioz,)

week of Jan 8:

- decide the work scope

week of Jan 15:

- meeting at Columbia, review LEXING code, transfer from Brian to Deniz

week of Jan 22:

- orientation to the project goals, LEXING within context, read onWordNet and ontologies

week of Jan 29:

- work out format of term extraction output from LKB with Hovy, Ambite, Philpot
- review the relations currently in SENSUS, determine which ones relevant to LEXING output and the requirements for SENSUS input
- Prototype 20 entries.

week of Feb 5:

- debug prototype entries based on input from ISI on API, requirements
- run larger test set, evaluate results
- finalize the workings out.

week of Feb 12:

- debug LEXING code; re-run 100 test items until agreement with our output and ISI input is met. First pass at delivering items.
- write evaluation software to randomly select from 6000 and others.

week of Feb 19:

- run alignments on 6000 EDC terms and deliver to ISI
- work with Raj to identify other glossary sources select test set from 3 new sources and run LEXING

week of Feb 26:

- debug, fix errors for EDC terms for ISI
- debug, fix errors for LEXING running on new sources

week of Mar 5:

- finalize alignments of 6000 EDC a test set of new terms
- teleconf with Hovy, Philpot, and anyone else from ISI to confirm that version 1 of the LEXING-SENSUS pipeline actually is flowing correctly.

week of Mar 12:

- identify errors from last week and fix them
- align new terms extracted from new sources. Output for SENSUS

week of Mar 19:

- collect data for merging (i.e. terms w/ >1 definition)
- align new terms extracted from new sources

week of Mar 26:

- merging research

week of Apr 2:

- data collection, structure, merging

week of Apr 9:

- debug and make sure the first demo works.

Terminology integration into SENSUS (ISI: Hovy, Ambite, Philpot)

week of Jan 8:

- update SENSUS to WordNet 1.6

week of Jan 15:

- update SENSUS to WordNet 1.6

week of Jan 22:

- finalize new SENSUS

week of Jan 29:

- run alignment tests on test terms
- work out format of term extraction output from LKB with JLK, Deniz

week of Feb 5:

- debug alignment routines based on tests

week of Feb 12:

- debug; run alignments on 6000 EDC terms

week of Feb 19:

- finalize alignments of 6000 (and new?) EDC (and other?) terms

week of Feb 26:

- model new sources

week of Mar 5:

- model new sources

week of Mar 12:

- align new terms extracted from new sources

week of Mar 19:

- align new terms extracted from new sources

Data sources

Addition of new data sources (ISI: Philpot, Ambite)

week of Feb 5:

- locate additional sources

week of Feb 12:

- locate additional sources

week of Feb 19:

- wrap new sources

week of Feb 26:

- wrap and integrate into SIMS new sources

week of Mar 5

- wrap and integrate into SIMS new sources

Semi-automated creation of domain models (ISI: Hovy, Ambite, Philpot)

week of Feb 5:

- outline process and heuristics
- create some hand examples

week of Feb 12:

- collaborate with data source wrapping to flesh out heuristics

week of Feb 19:

- start encoding heuristics

week of Feb 26:

- use heuristics to help with modeling of new sources

week of Mar 5:

- use heuristics to help with modeling of new sources
- refine heuristics and learning process

week of Mar 12:

- use heuristics to help with modeling of new sources

week of Mar 19:

- as time permits, continue with learning process and heuristics

week of Mar 26:

- as time permits, continue with learning process and heuristics

week of Apr 2:

- as time permits, continue with learning process and heuristics

In-memory Processing (Columbia: Ross)

week of Mar 26:

- integrate in-memory processing control scripts with system

week of Apr 2:

- complete system integration

week of Apr 9:

- test and debug

week of Apr 16:

- demo system in DC

Interface (Columbia: Feiner)

week of Jan 29:

- work out transfer of SENSUS to interface group with Steve

week of Feb 5:

- work out interaction between SIMS and interface

week of Mar 26:

- integrate with interface with rest of system

week of Apr 2:

- complete system integration

week of Apr 9:

- test and debug

week of Apr 16:

- demo system in DC

System (ISI: Philpot, Ambite, Hovy)

week of Jan 29:

- work out transfer of SENSUS to interface group with Steve

week of Feb 5:

- work out interaction between SIMS and interface
- work out format of term extraction output from LKB with JLK, Deniz

week of Feb 12:

- design new architecture to incorporate in-memory processing with Ken

week of Mar 26:

- integrate in-memory processing control scripts
- integrate with interface

week of Apr 2:

- complete system integration

week of Apr 9:

- test and debug

week of Apr 16:

- demo system in DC

3.2 Search for Additional Data

We have performed a search for additional data to include in the system.

OPEC

<http://www.opec.org/Publications/Publications.asp#ASB>

OPEC has an annual statistical report, latest is from 1999, for \$60. It looks a bit limited. The data are available electronically, but it's not clear you get it automatically when you buy the report. We have asked for a free copy as an educational institution.

IEA

<http://www.iea.org/stats/files/glance.htm>

International Energy Agency has a lot of data. Their most comprehensive data product is \$7500 a copy. We have asked for a demo copy, which they said was available. It looks as if most of the OECD data links are in fact redirects to IEA.

In addition,

<http://www.iea.org/stats/files/sel.htm>

is a small country-based web which we could wrap, but with very little data.

OECD

OECD does have some small monthly price statistics of mostly irrelevant quantities, in Excel and PDF versions.

API

<http://www.api.org/statistics/PDFFiles/paygas.pdf>

American Petroleum Institute has lots of data but most of it looks like it's targeted to petroleum industry executives with deep pockets. The above URL points to a file with propaganda about how energy taxes are too high.

UN

http://www.un.org/Depts/unsd/sd_economic.htm

<http://www.un.org/databases/index.html>

The UN's site is a mess. We did not find anything using either of these two seemingly good departure points.

US States

http://dir.yahoo.com/Science/Energy/Government_Agencies/United_States/

<http://www.tax.state.ak.us/FAQ/gasolineprices.htm>

Other states might have statewide agencies such as the CEC. In fact Alaska, Maryland, South Carolina, and Washington have sites listed in the below, but only California has price data. Alaska does republish some EIA data.

USC

USC has a list of electronic resources. Some of them are about economic data:

<http://www.usc.edu/isd/elecresources/subject/Business.html>

follow the links under the heading "Data sets/Statistics".

EIA: States and other (Philpot, Ambite, Hovy; ISI)

In the EIA state pages there is additional data in csv format that we could incorporate quite easily to our system. See

http://www.eia.doe.gov/emeu/states/main_ca.html

Also, Energy Consumption (PDF, ~90KB) or Data Files (Comma-delimited):

<http://www.eia.doe.gov/pub/state.data/data/CA.csv>

Energy Prices and Expenditures (PDF, ~95KB) or Data Files (Comma-delimited):

<http://www.eia.doe.gov/pub/state.prices/data/CA.csv>

Every state seems to have these two files that contain a lot of time series (203 per state).

So, we can get about 10000 more series.

Wrapper factory

The World Wide Web Wrapper Factory (W4F) is a pretty nice tool, but it's designed for HTML and they've gone commercial:

<http://db.cis.upenn.edu/W4F/>

<http://www.tropea-inc.com/>

—right now their demo version is not available, but see the examples at

<http://db.cis.upenn.edu/W4F/Examples/XML-Gateway/index.html>

3.3 System Integration (ISI and Columbia)

In the project meeting at Columbia in January 2001 we established contact with Ross and Feiner, who replaced Gravano and Hatzivassiloglou respectively. In regard to their respective tasks, ISI's role is to help design the overall system and to achieve full system integration.

Two types of user

Essentially, two kinds of users wish to access heterogeneous distributed data of the kind we are handling: expert analysts who work with data every day, and casual / one-time users who need to know something for an article they are writing or for a case that is being made.

These two types of user want different functions from the system. The experts want fast access to just the data they have decided is in their focus, and they want to be able to play around with various combinations of query settings. They are willing to wait overnight while the system prepares the data for them, because tomorrow their task still awaits. In contrast, one-time users want the most up-to-date data, and are not willing to wait overnight, but because they are not going to manipulate the data so intensely, do not need as instantaneous feedback either.

Thus far, we have focused on supporting the one-time users. The SIMS planner provides access to the most recent information contained in any of the online databases that have been incorporated. It does, however, have to plan the detailed data query from the user's request.

This year we are planning to add a major new functionality to the system. In order to support the needs of the expert analyst, we will include the in-memory processing modules developed at Columbia by Ken Ross. This capability is very fast and allows rapid recombination of query parameters, but cannot work with data that changes more than daily, since the processing module has to pre-compile index structures into the database.

Our overall plan is that Ross's in-memory processing, which is very fast but is limited to data that has been pre-indexed, will be used to show how the system can support users

like EIA analysts, who need to be able to combine data in lots of ways experimentally. Complementing this, the current system design continues as before to use SIMS to plan access to any database, no matter how recently the data came online. SIMS will decide when it is appropriate to route the user's query to the in-memory processing module and when to the SIMS planner.

In-memory processing (Columbia: Ross)

We have created the following integrated system design and are in communication with Ross to work out the details. Thus far, the plan is shaping up as follows:

- ISI will create one massive database that contains all the data we have currently incorporated into the system to date. The size will be between 250,000 and 300,000 lines.
- ISI will port this database to Columbia.
- Ross will compile his indexing and access indices into this database.
- ISI and Ross will develop an API under which queries can be passed from SIMS to the in-memory module of Ross.
- ISI and Ross and Feiner will work out the method by which the data can be displayed.

Interface (Columbia: Feiner)

Feiner is focusing on ontology-based menu creation and browsing at first. He has requested that a copy of SENSUS be deployed at Columbia, to facilitate development. We are working out the details. It is likely that Andrew Philpot from ISI will be given an account at Columbia and then deploy and take care of SENSUS there, for Feiner and this student to work with.

When the interface is ready, we will do whatever is required to make it work together with the rest of the system. This involves mutual development of:

- Query and data transfer APIs
- Conversion of menu queries into SQL
- Ontology query APIs

d. **User Testing** (Columbia: Bourne)

Walter Bourne from Columbia will design a series of tests and apply them to measure the usability of the interface and the system in general.

4. Work with Government Partners

4.1 EIA work

We have been in contact with Cal Kilgore from EIA, and have explained that the data they provide, while useful, is not dense enough in the data space to make interesting queries possible. Mr. Kilgore outlined his wishes for certain data to be reformatted and integrated. As described in Section 1.2 above, we then communicated with his subordinate Henry Weigel, and (using existing ISI expertise and technology) were able to

access, wrap, and reformulate over 60% of the problematic data in under a week.

With regard to the customized HTML reports we produced for the EIA (described in Section 1.2 above), here are the future activities planned:

- A. The particular data we extracted (Petroleum Marketing Monthly/Annual) are both consistent with and complementary to the other EIA data we have used (OGIRS and Petroleum Supply Monthly). Therefore, we can retarget this data into our current and ongoing Energy Data Collection application.
- B. At the moment, we are not using Ariadne to extract footnotes from the page footer subdocuments. This is only slightly more involved than the above. In EDC, we have a modeling convention wherein we use three model-level overlays on each modeled source, allowing us to associate an arbitrary number of independent footnotes to tables, columns, rows, and individual cells; this modeling framework can be applied directly to this data if we generate text wrappers for each form type's footnote reports. Currently, each footnoted measurement is generated as a tuple of [tag,value], but it would be straightforward to abstract this out into individual raw data, footnote tag, and footnote attachment reports (as we did for the EIA PSM and BLS web data for EDC).
- C. General geometric reasoning to understand multi-column headings (e.g., http://www.isi.edu/~philpot/eia/form1/monthly/tables31.txt_page_001_head) is not currently a part of Ariadne's reasoning process. There are approaches in the table understanding and induction community that we would like to investigate further.

4.2 Fedstats meeting in March

We are planning a meeting with the FedStats group, to show current progress and to elicit feedback and comments to guide our work. We had hoped to meet in late February, but will probably meet in March.

4.2 DG Workshop in April

We have established communication with Carol Hert in order to help her organize the DG workshop in DC in April. We have provided a list of discussion topics and offered to host a session.

6. dg.o 2001 Conference

This work was performed under a separate grant to ISI.

5.1 Hotel

After a site visit by Val Gregg and Chris Wingo, a contract was signed with the conference hotel, Crowne Plaza Redondo Beach. A deposit of \$35,000 has been paid to the hotel. For information about the hotel see

<http://www.basshotels.com/crowneplaza? franchisee=REDCP>.

5.2 System Demonstrations

We worked with BAE. A representative of BAE visited us in January 2001, for discussions about the hardware needs. He also visited the hotel with Susan Lapin. Things went smoothly at the conference.

5.3 Website

<http://www.dgrc.org/dg.o2001/> has the following structure:

- announcement
- call for presentations
- registration
- participants
- hotel
- about DGRC
- about dg.o

5.4 Conference Program

Overall plan:

	Monday	Tuesday	Wednesday		
7:15	Press breakfast				
8:00	Opening Keynote Address Stuart Lynn (Director of ICANN)	Special Session: Case Studies 1. COPLINK: Hsinchun Chen (U of Arizona), 2. Forestry Service Portal: Lois Delcambre (OGI), Tim Tolle (USDA Forest Service), Mathew Weaver (OGI)	Panel 5: Research Frontiers in DG Valerie Gregg (NSF, chair), Bob Chadduck (Nat Archives), Cathy Dippo (BLS), Peter Bloniarz (CTG and SUNY Albany)		
9:30	Break	Break	Break		
10:00	Papers 1: Handling Geospatial Data 1. Stefanidis et al. (U Maine), 2. Harms et al. (U Nebraska), 3. MacEachren et al. (Penn State), 4. Malyankar (Arizona State U), 5. Wojciechowski & Scott (Rice), 6. Samet et al. (U Maryland)	Panel 2: Digital Democracy Tom Temin (editor Gov Comp News, chair), Jane Fountain (Harvard U), Anthony Maddox (UCLA), Keith Thurston (GSA)	<table border="1"> <tr> <td> Panel 6: Technology Transfer Steve Cochran (CEG, chair), Robin Williams (IBM), Yigal Arens (ISI), Dina Lozofsky (USC), Mary Striegel (Nat'l Park Svce), Kevin Franklin (SDSC) </td> <td> Special session: Graduate Student Issues Roslin Hauck (U of Arizona, chair) </td> </tr> </table>	Panel 6: Technology Transfer Steve Cochran (CEG, chair), Robin Williams (IBM), Yigal Arens (ISI), Dina Lozofsky (USC), Mary Striegel (Nat'l Park Svce), Kevin Franklin (SDSC)	Special session: Graduate Student Issues Roslin Hauck (U of Arizona, chair)
Panel 6: Technology Transfer Steve Cochran (CEG, chair), Robin Williams (IBM), Yigal Arens (ISI), Dina Lozofsky (USC), Mary Striegel (Nat'l Park Svce), Kevin Franklin (SDSC)	Special session: Graduate Student Issues Roslin Hauck (U of Arizona, chair)				
11:30		Videos	Wrapup session		

12:00	Lunch IT of Ancient Documents Bruce Zuckerman, Marilyn Lundberg, Leta Hunt (USC)		Lunch	Lunch (NSF PIs)	Lunch
1:30	Panel 1: Internet Voting David Cheney (Internet Policy Institute, chair), David Jefferson (Compaq), Paul Herrnson (U of Maryland), Jane Fountain (Harvard U), David Elliott (State of Washington)		Panel 3: Issues in Privacy Sal Stolfo (Columbia U & iPrivacy.com, chair), Ari Schwartz (CDT), Blake Harris (editor GTN), Gene Tsudik (UC Irvine)		
3:00	Break				
3:30	Break				
4:30	Papers 2: Methods for Data Access (Peninsula) 1. Adam et al. (Rutgers et al.), 2. Bouguettaya et al. (Virginia Tech, Purdue), 3. Hovy et al. (USC/ISI and Columbia), 4. Ambite et al. (USC), 5. Zhang and Zhu (SUNY Buffalo), 6. Iyengar (Stanford U)	Papers 3: Interfaces for Data Collection and Display (Pacific) 1. Schober (New School U) and Conrad (BLS), 2. Shulman (Drake U), 3. Cheng et al. (U Maryland), 4. Marchionini et al. (UNC et al.), 5. Nusser et al. (Iowa State U et al.)	Panel 4: Opportunities in Government for DG Research Sharon Dawes (CTG, chair), Charlie Rothwell (CDC), Mark Bembem (DoD), Bob Maslyn (GSA)		
5:00	Break				
5:30	Birds of a Feather sessions				
6:30					
7:00	Demos and posters		Demos and posters		
9:00					

Papers 1: Handling Geospatial Data

Chair: Gary Marchionini

- A. Stefanidis, P. Partsinevelos, P. Agouris (University of Maine). *Using Lifelines for Spatiotemporal Summaries.*
- S.K. Harms (University of Missouri-Columbia), S.E. Reichenbach (University of Nebraska-Lincoln), T. Tadesse (University of Nebraska), W.J. Waltman (University of Nebraska). *Data Mining in a Geospatial Support System for Drought Risk Management.*
- A.M. MacEachren, M. Wheeler, F. Hardisty, M. Gahegan, X. Dai, D.-S. Guo, M. Takatsuka (Penn State University). *Supporting Visual Integration and Analysis of*

Geospatially-Referenced Statistics through Web-Deployable, Cross-Platform Tools.

- R. Malyankar (Arizona State University). *Maritime Information Markup and Use in Passage Planning.*
- W.C. Wojciechowski and D.W. Scott (Rice University). *Conditioning Multiple Maps.*
- H. Samet, F. Brabec, G. Hjaltason (University of Maryland, College Park). *Interfacing the SAND Spatial Browser with FedStats Data.*

Papers 2: Methods for Data Access

Chair: Jamie Callan

- N.R. Adam (Rutgers University), F. Artigas (Rutgers), V. Atluri (Rutgers), S.A. Chun (Rutgers), S. Colbert (NL Office of IT), M. Degaratu (Columbia University), A. Elbeid (NJ Office of IT), V. Hatzivassiloglou (Columbia), R. Holowczak (City University of NY), O. Marcopolus (NJ Office of IT), P. Mazzoleni (University of Milan), W. Rayner (CIO, State of NJ). *E-Government: Human Centered Systems for Business Services.*
- Bouguettaya (Virginia Tech), A. Elmagarmid (Purdue University), B. Medjahed (Virginia Tech), M. Ouzzani (Virginia Tech). *An Ontology-based Infrastructure for Government Databases.*
- E.H. Hovy (USC Information Sciences Institute), A. Philpot (Information Sciences Institute), J.L. Ambite (Information Sciences Institute), Y. Arens (Information Sciences Institute), J.L. Klavans (Columbia University), W. Bourne (Columbia University), D. Saros (Columbia University). *Data Acquisition and Integration in the DGRC's Energy Data Collection Project.*
- J.L. Ambite (USC Information Sciences Institute), C. Shahabi (University of Southern California), R.R. Schmidt (University of Southern California). *Fast Approximate Evaluation of OLAP Queries for Integrated Statistical Data.*
- Zhang and L. Zhu (SUNY Buffalo). *Metadata Generation and Retrieval of Geographic Imagery.*
- S. Iyengar (Stanford University). *Who Needs the Media?*

Papers 3: Interfaces for Data Collection and Display

Chair: Peggy Agouris

- M.F. Schober (New School University) and F.G. Conrad (Bureau of Labor Statistics). *Adaptive Interfaces for Collecting Survey Data from Users.*
- S.W. Shulman (Drake University). *Citizen Agenda Setting: The Electronic Collection and Synthesis of Public Commentary in the Regulatory Rulemaking Process.*
- W.C. Cheng, C.-F. Chou, L. Golubchik, S. Khuller, H. Samet (University of Maryland). *Scalable Data Collection for Internet-based Digital Government Applications.*

- G. Marchionini (University of North Carolina), C. Hert (Syracuse University), B. Shneiderman (University of Maryland), L. Liddy (Syracuse University). *E-Tables: Non-Specialist Use and Understanding of Statistical Data*.
- S. Nusser (Iowa State University), K. Clarke (University of California, Santa Barbara), M. Goodchild (University of California, Santa Barbara), L. Miller (Iowa State University). A New Framework for Computer-Assisted Data Collection.

Posters

We had 12 posters.

Monday evening:

Sherri Harms	University of Missouri & University of Nebraska	Discovering Associations between Climatic and Environmental Variables for Supporting Drought Decision Making
Dan Carr	George Mason University	Selected Designs for Communicating Federal Statistical Summaries
Ram Chellappa	University of Southern California	Contrasting Scientific Assessment of Privacy with Perceived Privacy: Implications for Public Policy
Jamie Callan	University of Massachusetts and Carnegie Mellon University	A Language-Modelling Approach to Metadata for Cross-Database Linkage and Search
Scott Midkiff	Virginia Tech	Rapidly Deployable Broadband Wireless Communicators for Emergency Management
William Wojciechowski, David Scott	Rice University	Conditioning Multiple Maps

Tuesday evening:

Aidong Zhang	SUNY Buffalo	Keyblock-based Approach for Geographic Image Retrieval
Stuart Shulman	Drake University	Citizen Agenda Setting: The Electronic Collection and Synthesis of Public Commentary in the Regulatory Rulemaking Process
Bill McIver	Brown University	Integrating Critical Theory into Studies of the Citizen-Government Digital Divide
Fred Conrad, Michael Schober	BLS and New School University	Adaptive Interfaces for Collecting Survey Data from Users
Leana Golubchik	University of Maryland at College Park	Scalable Data Collection for Internet-based Digital Government Applications
Seung-Yong Rho	Rutgers	Citizens' Trust in Digital Government: Toward Citizen Relation Management

System demos

We had 37 demos (one of which is a poster with a table).

Monday demos:

Steven Feiner and Surabhan Temiyabutr	DGRC, Columbia University	Exploratory Design for a Database Query Interface
Eduard Hovy, Andrew Philpot, and Jose Luis Ambite	DGRC, USC/ISI	Modeling in Ontology-Based Access to Multiple Heterogeneous Databases
Judith Klavans	DGRC, Columbia University	Glossary Mining in the Energy Data Collection Project
Ken Ross	DGRC, Columbia University	The DataCube
Jose Luis Ambite, Cyrus Shahabi and Rolfe Schmidt	DGRC, USC/ISI and USC	Fast Approximate Evaluation of OLAP Queries for Integrated Statistical Data
Chris Wingo	Council for Excellence in Government	eGov Blueprint: Council for Excellence in Government
Raphael Malyankar	Arizona State University	Maritime Information Markup and Use in Passage Planning
Sarah Nusser and Peisheng Zhao	Iowa State University	A Framework and Testbed for Data Collection in a Mobile Field Environment
Leslie Miller, Nikhil Sathe, and Hua Ming	Iowa State University	An Infrastructure for Delivering Geospatial Data from Heterogeneous Data Sources to the Field
Michael Goodchild	UC Santa Barbara	Using Geolibraries in the Field
Keith Clarke and Andrea Nuerenberger	UC Santa Barbara	A Prototype Wearable System for Field Computing
Kincho Law	Stanford University	Information Infrastructure for Regulation Management Compliance Checking
Shinto Iyengar	Stanford University	Design and Use of Campaign Handbooks
Alan MacEachren	Penn State University	GeoVITSA Studio: Supporting visual integration and analysis of geospatially-referenced data through web-deployable, cross-platform tools
Lois Delcambre, Mathew Weaver, Shawn Bowers	Oregon Graduate Institute	Harvesting Information to Sustain our Forests
Steve Minton, Brian Pelz	Fetch Technologies	Internet Application Integration (the Fetch Platform)
Hanan Samet	University of Maryland at College Park	Interfacing the SAND Spatial Browser with FedStats Data
Rosie Hauck	University of Arizona	COPLINK
Michael Chau and Homa Atabakhsh	University of Arizona	COPLINK: Building an Infrastructure for Law Enforcement Information Sharing and Collaboration: Design Issues and Challenges

Tuesday demos:

Neill Scott	Stanford University	Improving Access for Blind and Deaf Computer Users: New developments in the Total Access System
Margaret Marks and Bill LaPlante	Census	Assistive Technology at the National Processing Center
Kelsey Rideout	BAE Systems	Making Commercial Products Accessible: A Case Study
John O'Looney	University of Georgia	Personalization Technology for Government Internet Services
Alan Karr	National Institute of Statistical Sciences	Web Systems that Disseminate Information but Protect Confidential Data
Soon Ae Chun and Nabil Adam	CIMIC, Rutgers University	Human Centered Systems for E-Government Business Services
Athman Bouguettaya	Virginia Tech	A Web-based Infrastructure for Government Databases
Christine Meers	Department of Transportation	Department of Transportation's Docket Management System
Patricia Cruse	California Digital Library, Office of the President	Counting California: a Gateway Integrated to Government Data from the California Digital Library
Erin Shaw, Lewis Johnson	USC/ISI	CARTE: Distance Education
Ilya Zaslavsky	SDSC	Mediation of XML Sources
Chaitan Baru	SDSC	The Sociology Workbench and DDI
Gary Marchionini	University of North Carolina	E-Tables: Non-Specialist Use and Understanding of Statistical Data
Rachel Taylor and Marshall DeBerry	Census and Department of Justice	The FedStats Website
Tony Stefanidis, Panayotis Partsinevalos, Peggy Agouris	University of Maine	Using Lifelines for Spatiotemporal Summaries
Robert Neches	USC/ISI	GeoWorlds Geospatial Information Management System
Chandrashekar Ramanan, Milind Tambe	USC/ISI	Adjustable Autonomy in Personal Assistants: An Illustration using the Electric Elves System
J. Etchemendy and D. Barker-Plummer	Stanford University	Heterogeneous Tools for Rationale Capture in Survey Instrument Design

5.5 Interviews

Learning from dg.o 2000, we have decided to videotape only interviews with speakers, and not the actual presentations. We sent questions to speakers before the conference so that they could prepare their comments.

5.6 Report

The conference report is being prepared in collaboration with RLA Associates. We have the design for the cover, the overall layout, and the content plan already done. At the time of writing we are collecting the final materials from participants (papers, slides, etc.), for inclusion. We aim to have the report completed by August 31, 2001.

6. dg Online Newsletter

The first issue of the dg.o Newsletter appeared in April 2001. The next issue is in preparation, due out July 1. Please see <http://www.dgrc.org/dg-online/>.

7. Biodiversity Polyclave

After Eduard Hovy attended the Biodiversity Initiative workshop organized by the NSF, we built a prototype polyclave that might be used as example to show collaboration between Computer Scientists and Biologists. It took approx one week to build.

The polyclave, which was presented to Jim Quinn, Tom Schnase, and others from the workshop, can be seen at <http://brawn.isi.edu:8888/>.

8. Publications

We have created a website with all the DGRC publications, as a central repository:
<http://www.dgrc.org/pubs/main.html> and <http://www.cs.columbia.digov/>.

Publications and Activities of EDC Project Digital Government Research Center

Technical papers

- Paper in volume, edited by W. McIver, 2001.
- Paper at AFCEA Database Colloquium, June 2001.
- Paper at dg.o 2001 Conference, May 2001.
- Overview article in IEEE Computer, Feb 2001
- Paper at Joint Statistical Conference, Aug 2000

Excerpts from proposals

- Proposal to NSF, 2000
- Proposal to NSF, 1999

Reports

- EDC Project Annual Report to NSF, 2000

Slides from Presentations

- EDC overview presentation, May 2000 (in PowerPoint)
- Columbia site visit presentation, Mar 2000 (in PowerPoint)

Conferences and Workshops

- dg.o 2001 Conference, May 2001: dg.o 2001
- dg.o 2000 Workshop, May 2000: dg.o 2000