Panel:
Review of Information Technology

Eduard Hovy, USC/ISI
Gio Wiederhold, Stanford (emeritus)
Ed Fox, Virginia Tech
Claire Cardie, Cornell
Big problem: accessing distributed info

Problem:
1. Rulemakers need accurate information about the domains their rules might affect (complex legal interactions, Gov agency operations, citizens’ lives, etc.)
2. Agencies and citizens need access to rules, often across several agencies’ purviews
3. Rules change; agencies change; people’s situations differ

(Part of) the solution: portals to live Government information

- Case study: the Energy Data Collection, and other portals
- Challenges for IT research
E-rulemakers might benefit from knowing what the world is like...some DG projects

- **Energy Data Collection (EDC) portal for access to distributed heterogeneous gasoline price data (Hovy, Klavans et al.; ISI and Columbia)**

Enable government workers and citizens to get access to gasoline price info collected by EIA, Census, BLS, CEC that is on CDs, websites, etc. Problem of nomenclature standardization and data incommensurability. **Technology**: database wrappers and domain models under large ontology.

- **Database middleware for distributed ontologies in state and federal family & social services (Bouguettaya, Medjahed, et al.; Virginia Tech and Purdue)**

Enable Indiana citizens and social workers to get timely information and services from local, state, and federal governments. Typically, a single mother has to deal with the Division of Family and Children, Bureau of Child Support, Medicaid Policy and Planning, Social Security, etc.: register for Medicaid, get food stamps, interact with social workers for child care, etc. The project allows mothers, social workers, government officials to organize, manage, and access welfare databases and services. **Technology**: database interconnections via ontology.
More portals…

• **Using controlled vocabularies as a knowledge base for natural resource managers (Delcambre, Tolle et al.; Oregon Graduate Institute and USDA Forest Service)**

Capture and provide forestry information for foresters, Gov agencies, and public, in order to counterbalance decentralization of Gov agency (and rulemakers) as well as support specific knowledge requests from public and others. Focus on 26 domains (wildlife, climate, vegetation, location, etc.). **Technology**: extract terminology by interview and from texts, create controlled vocabulary taxonomy, interrelate. The access website enables search and retrieval of useful documents, maps, etc.

• **Dynamic composition of workflows for customized eGovernment service delivery (Adam, Atluri, et al.; Rutgers)**

Create New Jersey Business Portal, website that enables business persons, entrepreneurs, and other citizens to understand and work with interrelated Gov services. Domain: when starting a new business, entrepreneurs need forms, information, etc., from numerous agencies. System aims to dynamically compose flowcharts to deliver customized plan of action. **Technology**: taxonomized collection of rules (ontology) about regulatory, financial, functional, temporal, etc. requirements; upon request, rules interact to assemble flowchart.
And two more

- REGNET: a formal infrastructure to support access to government regulations (Law, Wiederhold, Kerrigan, et al.; Stanford)
  Help people discover whether their enterprises conform to regulations about Environment, Disabled Access, etc.; suggest things they might have overlooked, etc. Technology: convert regulations into XML format, create taxonomy, add concepts, legal and logical interpretation annotations, find interrelationships among regulations.

- Helping EPA officials and their counterparts in Europe standardize terminology using the CARAT ontology (Hovy, ISI)
  Problem of inconsistent terminology and definitions in writing EPA pollution regulations for US, GB, Finland, Mexico, Germany, Holland…. EPA’s CARAT ontology is a starting point for ongoing discussions. Problem: align texts from other countries into CARAT for easier comparisons. Technology: text processing for entity extraction and automated ontology alignment.
Ex.: the heart of the EDC system

SIMS data access planner:
- SIMS tools ‘wrap’ data sources—build access routines using standardized domain model terms
- SIMS query access planner—SQL converted to access routines
- SIMS assembles results

SENSUS Ontology:
- Large term taxonomy to ensure standardization
- Domain model terms linked into it
- New terms scraped off website glossaries

Interface: built by partners (Columbia University)

- SENSUS ontology: 90,000 nodes
- Domain model: 500 nodes (built manually; for SIMS)
- LKB: 6000 nodes (NL term/info extraction from glossaries)
- Databases: 58,000+ time series (EIA OGIRS and others)
- Webpages: 60+ (BLS, CEC tables)
A general procedure

In the rulemaking process, if you find yourself doing something over and over
...you want a computer to do it. So:

• Step 0: find a computer scientist to work with
• Step 1: define the inputs
• Step 2: define the outputs
• Step 3: create a lot of examples
• Step 4: train a program to do the work:
  – Learning stage: recognize the pertinent features
  – Application stage: apply appropriate rules
• Step 5: evaluate and fix
A challenge problem

Help lawmakers and writers
- Collect raw data
- Make facts comparable

Help writers and citizens
- Collect citizen commentary
- Cluster by genre, opinion, pressure group

Help lawyers & archivists
- Collect newspaper and other comments and stories

SuperFinder:
- Access databases, info
- Analyze and display

SuperBboard:
- Cluster commentary
- Analyze and display

SuperAnalyst:
- Collect pertinent info
- Thread subtopics
- Analyze and store
Thank you!
Discussion?

Gio Wiederhold, Ed Fox, Claire Cardie, Ed Hovy
Case study: EDC

(ISI and Columbia)

- **Problem**: US Government has thousands of databases in over seventy FedStats agencies:
  - data is duplicated and near-duplicated,
  - even Government officials cannot find one another’s data!

- **Research challenge**: Provide access to many databases, for both sophisticated and casual users, in an easy-to-use and easy-to-understand format, without distorting data

- **Solution**: Create systems to provide easy, fast, and/or standardized access:
  - need method of ‘standardizing’ many databases
  - need multi-database access engine
  - need powerful user interface
SENSUS Browser

http://mozart.isi.edu:8003/sensus/sensus_frame.html
The Semantic Web

• The vision…
  system of annotations for webpages: standardized, Interlingual, semantic, cover all media, easily created by users, early rendered elsewhere…

• But this is not so easy! Need:
  – standardized interlingual semantic annotation termset, extensible (by whom?)
  – interpreter engines for text, pictures, numbers, graphs
  – Generator engines for all languages
Some existing termsets

- Classification Society of North America [http://www.pitt.edu/~csna/](http://www.pitt.edu/~csna/)
- Archives (National Union Catalog of Manuscript Collections (NUCMC) info in Library of Congress, covering more than 90,000 collections; National Inventory of Documentary Sources in the US (NIDS) names and detailed subject indexing of over 54,000 collections): [http://archives.chadwyck.com/infopage/moreinfo.htm](http://archives.chadwyck.com/infopage/moreinfo.htm)
- Numerous data documentation and other standardization efforts
- Several ontology standardization efforts (ANSI effort, SUO)
Learning terms, relations, rules

- Databases: automating database mappings (Halevy, Dominguez et al.; Mitra & Wiederhold; Rahn & Bernstein; Zaslavsky & Baru; many more…)
- Databases: data mining (Tukey; Hoaglin et al.; Fayyad and Uthurusamy…)
- Ontologies: automating mappings, learning content (Hovy et al.; Noy & Musen; Studer et al.; Agirre et al.…)
- Text mining: extracting terms etc. from text (Hearst; Dagan; Feldman; Manning…)

IT capabilities & research challenges

• Capabilities:
  – Various ways to handle database access
  – Various ways to index and handle textual info
  – Some ways to index and handle visual/audio info

• Challenges:
  – Automating ‘registration’ of information (databases, etc.): creating metadata / domain models and access methods
  – Supporting (automating?) terminology management: creating, extracting, disambiguating, cross-relating
  – Creating reasoning engines: extracting from text and databases more sophisticated reasoning rules for cross-reference, cross-validation, user guidance, etc.