

**CANIS**

## *The Interspace Prototype: An Analysis Environment for Semantic Interoperability*

*Bruce Schatz (PI), Kevin Powell, University of Illinois**Hsinchun Chen (coPI), Marshall Ramsey, Kris Tolle, University of Arizona*

### ABSTRACT

We are developing semantic indexing techniques that can be effectively computed for large-scale real-world collections. These indexing techniques are being incorporated into integrated prototypes of analysis environments. These prototype environments enable information analysts to easily navigate across subjects and across media to similar objects within and across collections. The prototypes incorporate scalable semantics, indexing on large collections that enables federation across repositories. The *Interspace* Prototype supports practical semantic interoperability, by enabling users to navigate across logical spaces of concepts, which are indexed repositories of physical collections. The prototypes have been demonstrated to be general enough to index real collections of millions of objects, both text and image. Studies have shown them useful enough to support classification comparable in quality to human indexers. Initial technology transfer of these prototypes has been effected into DoD analyst sites.

### SEMANTIC INDEXING

The semantic indexing techniques are based upon the context of units within the collection of objects. We have developed generic indexing technology for both text and image collections.

For text documents, the units are terms within documents. We have developed a generic noun phrase extractor and successfully parsed documents from a wide variety of sources. The context is co-occurrence frequency of which phrases occur with which phrases within documents in the collection. We have developed generic technology, which has uniformly processed a wide variety of sources to produce concept spaces. We have also developed category map technology for clustering similar documents within a collection — useful for navigation, whereas concept spaces cluster similar terms and are useful for search. These are based on statistical neural net techniques from self-organizing maps.

These sources have mostly been bibliographic databases of scientific papers, since these enable comprehensive coverage across subject areas in a discipline. For example, we indexed 4M abstracts from INSPEC (electrical engineering and computer science) and COMPENDEX (engineering) and 40M abstracts from MEDLINE (medicine). The MEDLINE run produced 280M noun phrases. These computations were performed on a parallel shared-memory supercomputer at NCSA. The disciplines are partitioned into community repositories, about 10K documents each, and indexes generated for each community. We now have an implementation for our integrated environment, which routinely computes semantic indexes on our parallel workstation cluster at CANIS (our systems research laboratory at Illinois).

For image collections, we have concentrated on aerial photographs, which are monochrome natural scenes. The same indexing technology as for text was used to compute concept spaces and category maps, with similarly good results for large real-world collections. Thus, we have successfully developed generic semantic indexing technology, which performs well across subjects and across media. When the objects are aerial photographs, the units are texture tiles. We have processed 800 photos of Southern California from the Map and Imagery Library at the University of California at Santa Barbara, each about 25 MB. This produced 6M texture tiles from 200K discrete regions. To correlate text with image, we also indexed 1M bibliographic abstracts from Petroleum Abstracts, GEOREF, and COMPENDEX. A third media type, number, is provided by NASA AVHRR vegetation records for 8000 places.



## ANALYSIS ENVIRONMENTS

The Interspace Prototype is an integrated analysis environment that supports semantic interoperability across subjects and across media. In this year's demonstrations, the text prototype and the image prototype are each separately integrated across different index types.

The Interspace Prototype (text) is an object-oriented analysis environment implemented in Smalltalk and Versant. It uniformly processes noun phrases on text documents and generates concept spaces and category maps from these for each community repository. Its functionality provides concept navigation within and across repositories. Within a repository, a user can transparently navigate from similar documents within a cluster; to phrases within a document; to related phrases within the collection; to documents containing a selected related phrase. Desired documents can be retrieved without explicit searching or knowing the specific terminology used within the collection. Sample communities from electrical engineering (INSPEC) and from medicine (MEDLINE) are available with a full bank of semantic indexes.

Concept switching is the paradigm for semantic interoperability in the Interspace. The user can transparently navigate at the concept level from one space (community repository with semantic index) to another space. We have experimented extensively with concept switching, supported by syntactic transformations. These are general for all subject domains, e.g. affecting word order (exchanges, deletions) or word conflation (stemming, variants). We have also experimented with semantic transformations, specifically for the domain of medicine. The National Library of Medicine UMLS (Unified Medical Language System) has specific knowledge bases useful to implementing semantic transformations. We will be demonstrating concept switching with term variations and term transformations using the UMLS Semantic Map and Metathesaurus.

The Interspace Prototype (image) is currently a separate subsystem. It uses the spatial coordinates, which are uniform across different media sources, to provide semantic interoperability specific to geographical knowledge. Concept switching within the image collection of aerial photographs and within the text collection of bibliographic abstracts is supported by concept navigation using concept spaces and category maps on the respective units and objects. Concept switching between image and text is supported by the Geographical Names Information System from the U.S.

Geological Survey, a useful knowledge source for identifying relationships between precise coordinates and fuzzy place names. The Interspace GKRS (Geographical Knowledge Representation System) contains 1.8 million place names, organized hierarchically into 15 geographic feature classes. The GKRS system allows users to retrieve multimedia geoscience content with fuzzy queries, e.g. "Find me information, in the form of texts, vegetation records, and images, about orchards along the Santa Cruz River in Southern California".

The screenshot displays a medical abstract interface. On the left is a text abstract titled "Nonsteroidal anti-inflammatory drugs, arthritis, and gastrointestinal bleeding in elderly in-patients." The abstract text discusses a study of 531 consecutive admissions to an acute geriatric unit in England, noting a relationship between NSAID use and upper gastrointestinal bleeding. A callout bubble points to the text "simple analgesics" with the text "click here to see concept for: simple analgesics". On the right, there are two "Related Terms" boxes. The top box is for "TENIDAP" and lists related terms like "MUCOPOLYSACCHARIDURIC ACID INHIBITORS", "DYSPEPSIA", "GASTROINTESTINAL BLEEDING", "GASTROINTESTINAL ULCERATION", "GASTROINTESTINAL PERFORATION", "GASTROINTESTINAL OBSTRUCTION", "GASTROINTESTINAL ISCHEMIA", "GASTROINTESTINAL DYSMOTILITY", "GASTROINTESTINAL INFLAMMATION", "GASTROINTESTINAL STENOSIS", "GASTROINTESTINAL DIVERTICULITIS", "GASTROINTESTINAL ANGIOBLASTOMA", "GASTROINTESTINAL POLYPOSIS", "GASTROINTESTINAL NEOPLASIA", "GASTROINTESTINAL TUMOR", "GASTROINTESTINAL CYST", "GASTROINTESTINAL ABSCESS", "GASTROINTESTINAL HEMATOMA", "GASTROINTESTINAL HEMORRHAGE", "GASTROINTESTINAL PERITONITIS", "GASTROINTESTINAL ABSCESS", "GASTROINTESTINAL HEMATOMA", "GASTROINTESTINAL HEMORRHAGE", "GASTROINTESTINAL PERITONITIS". The bottom box is for "SIMPLE ANALGESIC" and lists related terms like "MAINTENANCE THERAPY", "CLASS II", "ANTI-INFLAMMATORY AGENT", "MAGNESIUM", "GASTROINTESTINAL BLEEDING", "GASTROINTESTINAL ULCERATION", "GASTROINTESTINAL PERFORATION", "GASTROINTESTINAL OBSTRUCTION", "GASTROINTESTINAL ISCHEMIA", "GASTROINTESTINAL DYSMOTILITY", "GASTROINTESTINAL INFLAMMATION", "GASTROINTESTINAL STENOSIS", "GASTROINTESTINAL DIVERTICULITIS", "GASTROINTESTINAL NEOPLASIA", "GASTROINTESTINAL TUMOR", "GASTROINTESTINAL CYST", "GASTROINTESTINAL ABSCESS", "GASTROINTESTINAL HEMATOMA", "GASTROINTESTINAL HEMORRHAGE", "GASTROINTESTINAL PERITONITIS".

## TECHNOLOGY TRANSFER

Initial technology transfer from the Interspace Prototype has been demonstrated at DoD analyst sites. Some of our linguistic processing techniques has been applied to the Intelink project at the National Security Agency. A subsystem called Enhanced Phrase Search (EPS) was incorporated into Intelink for high-precision phrase-based searches in May 1999. This was the follow-up to detailed technical visits to our project.

The integrated Interspace Prototype was specifically demonstrated for the VIC (Virtual Information Center) group at PACOM (formerly called PMAD). These analysts search public domain web sources to prepare briefing backgrounds for Pacific Command. We generated semantic indexes from the "Asian economics" section of ABI/Inform, a bibliographic database of manufacturing information about companies. A targeted demonstration of concept navigation within the Asian Economics community repository was given for VIC/PACOM by the IM/ITO technology transfer group from BBN in January 1999.

For further information, please contact:  
Bruce Schatz, schatz@canis.uiuc.edu  
CANIS Laboratory, University of Illinois  
704 S. Sixth St., Champaign, IL 61820  
217-244-0651 fax 217-333-6869