SAND2007-2039C

The Critical First Year: Introducing Semantic Technologies into an Organization

John M. Linebarger, PhD

Bettina K. Schimanski, PhD

Sandia National Laboratories

23 May 2007

Los Alamos

Sandia National Laboratories

Outline

- Technology Diffusion is a Process of Organizational Change
- The Critical First Year
  - Timeline
  - Semantic Technology Projects
- Reflections and Recommendations
  - Organizational Issues
  - Marketing Issues
  - Ontology Development Issues
  - Tools and Support Issues
- Resources

Diffusion of Innovations

- Everett M. Rogers (University of New Mexico)
- Technology diffusion is a process of organizational change
- The change process follows a predictable pattern
- Early adopters (opinion leaders) are vitally important in the process

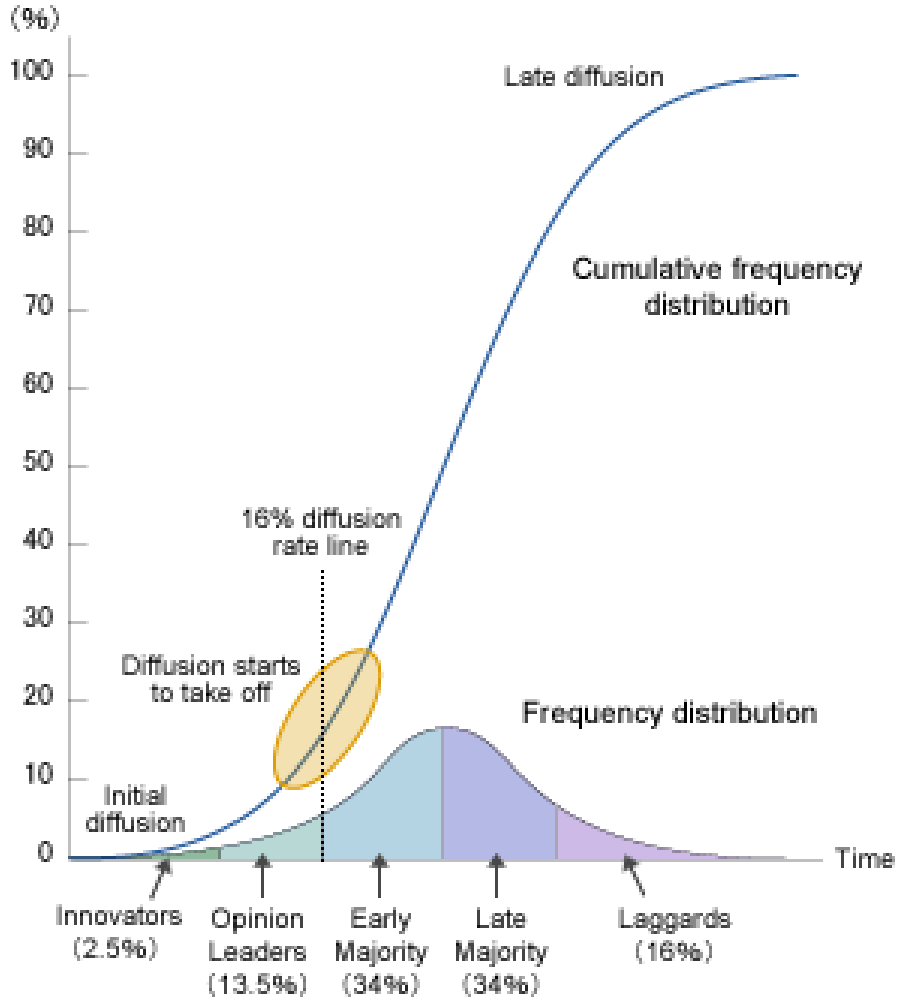


Image source: http://www.mitsue.co.jp/english/case/concept/02.html

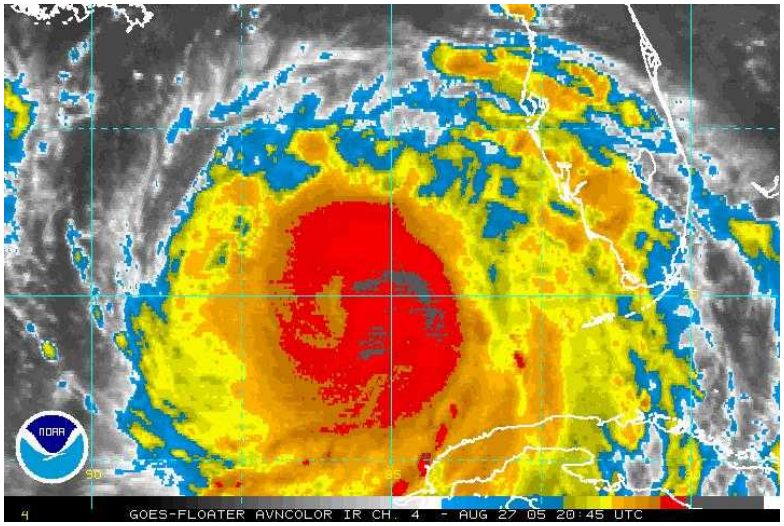
Timeline of "The Critical First Year"

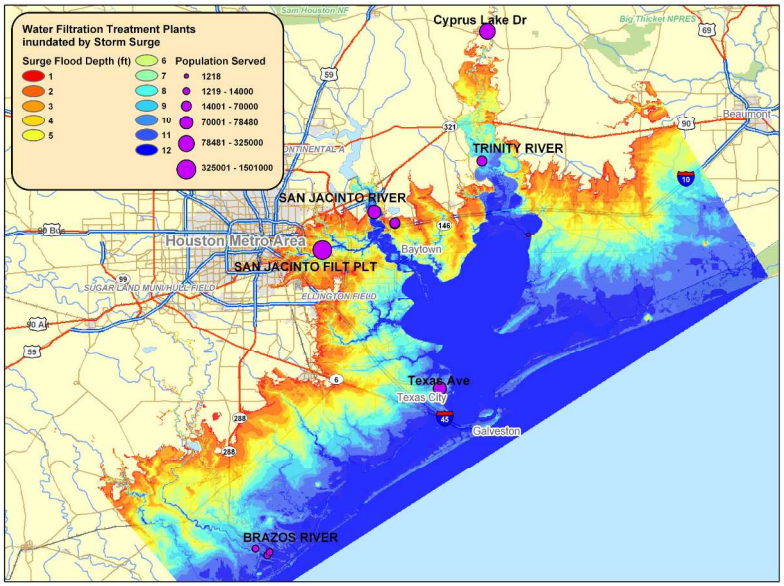
- Nov. 2005—Full-time on Semantic Technologies for the NISAC program
- Dec. 2005—Semantic Working Group formed
- May 2006—Pædogogical Application created
- July 2006—Semantic Navigation prototype and user tests with the help of student interns
- Oct. 2006—Lockheed Martin Shared Vision project involvement; hired a second person for Semantics
- Nov. 2006—FEMA IPAWS project involvement
- Dec. 2006—First Semantic Web application placed into production
- Jan. 2007—Virtual Manufacturing project involvement

The NISAC Program

- The National Infrastructure Simulation and Analysis Center (NISAC) program is sponsored by the Department of Homeland Security (DHS)

- NISAC is often called upon to quickly analyze the impact on critical infrastructures of a potential future event
  - Fast Analysis and Simulation Team (FAST) exercises
  - Time-limited (from four hours to several days)

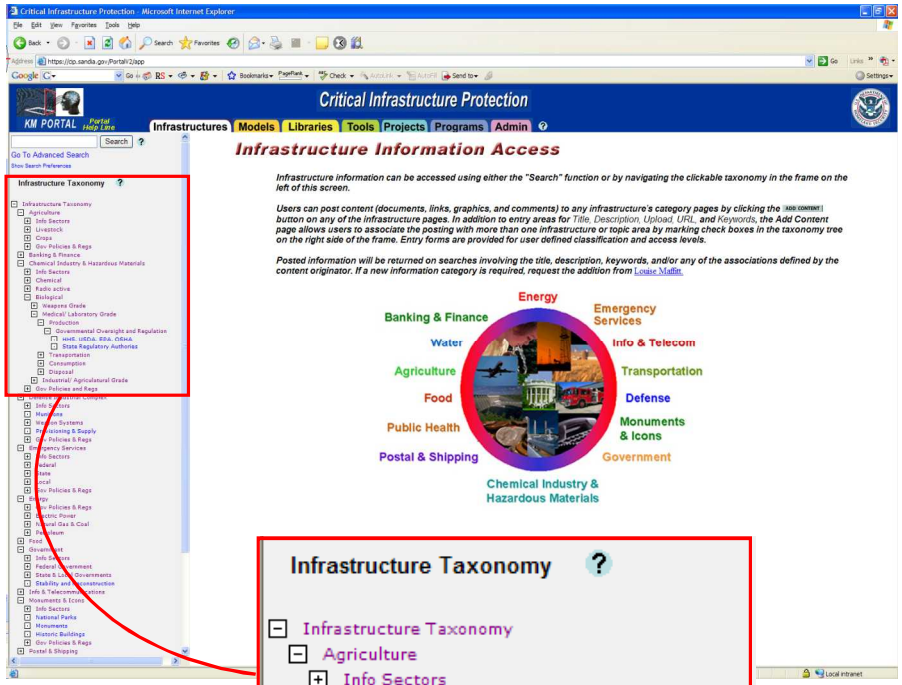




Los Alamos

Sandia National Laboratories

NISAC CIP KM Portal



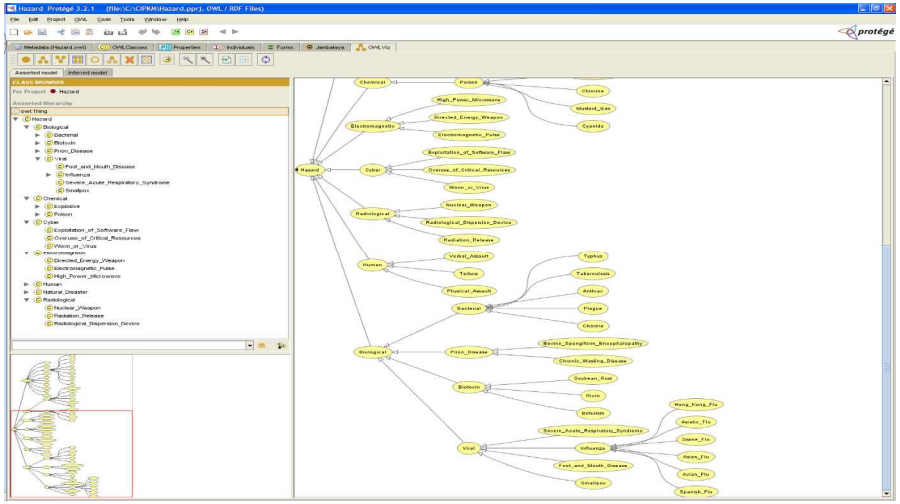
- Critical Infrastructure Protection (CIP) Knowledge Management (KM) Portal
- Supports rapid access of information during a FAST exercise
  - Documents
  - Presentations
  - Media files
  - Links to external Web pages
- Organization of information using multiple taxonomies has not proven to be sufficient
  - Programs
  - Projects
  - Infrastructures
  - Tools
  - Models
- Keyword search has well-known limitations

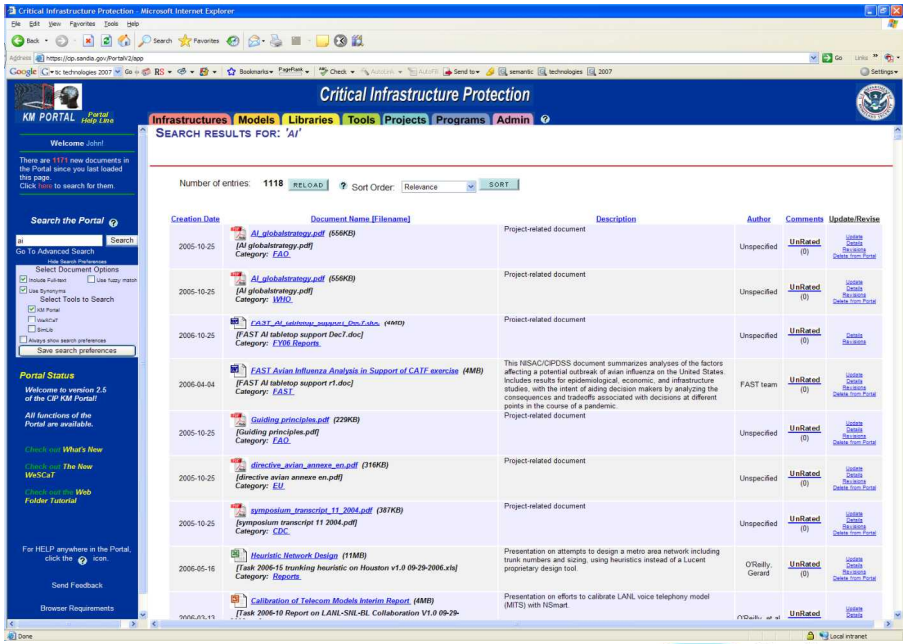
Los Alamos

Sandia National Laboratories

Semantics in the NISAC Program

- Ontology development
  - Critical Infrastructure Protection (CIP)
  - CIP Knowledge Management (KM)
- Semantic Navigation of the CIP KM Portal (prototype)
- Synonym expansion of keywords used to search the CIP KM Portal (production)



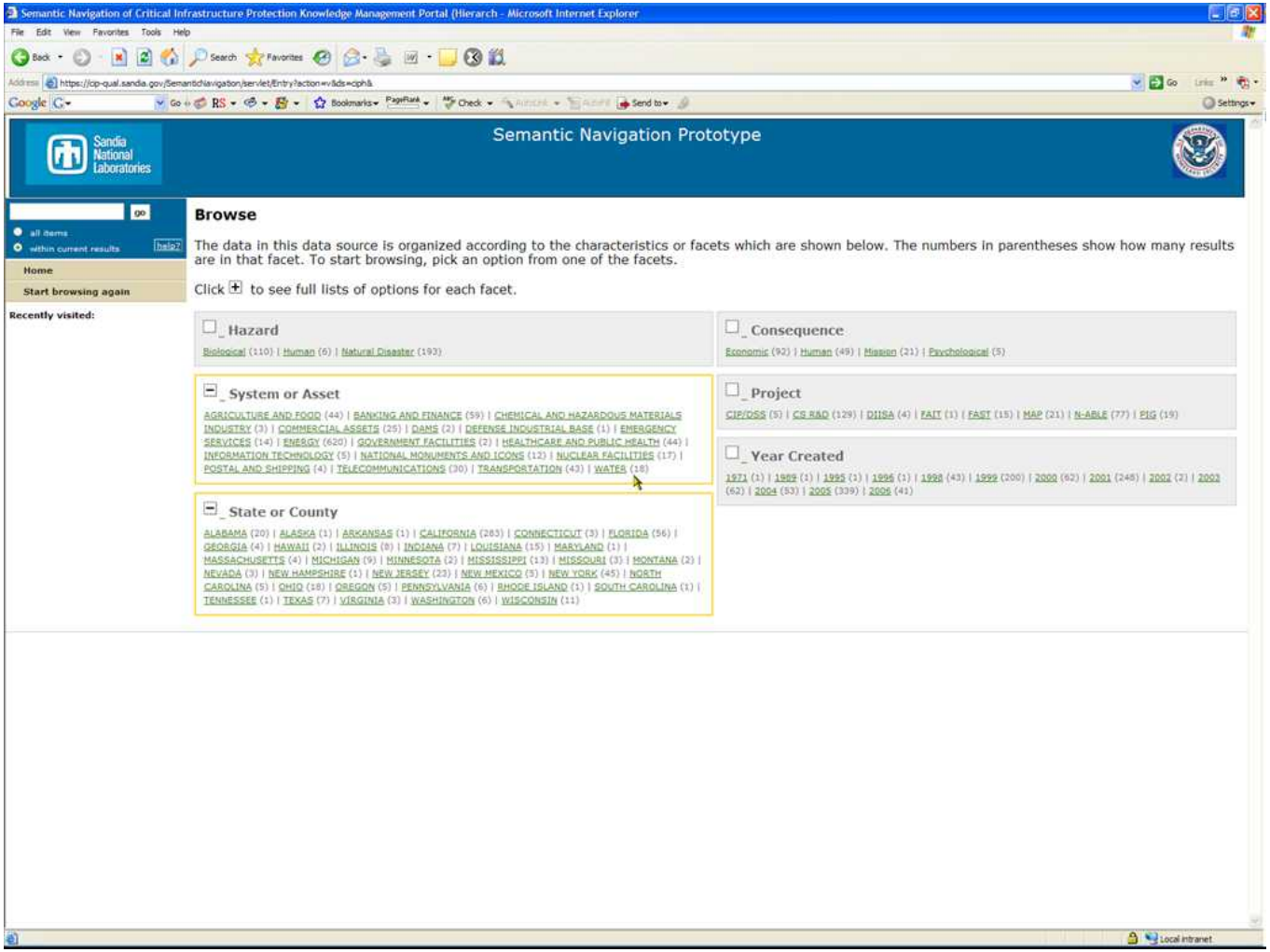


NISAC

Los Alamos

Sandia National Laboratories

Semantic Navigation of the CIP KM Portal



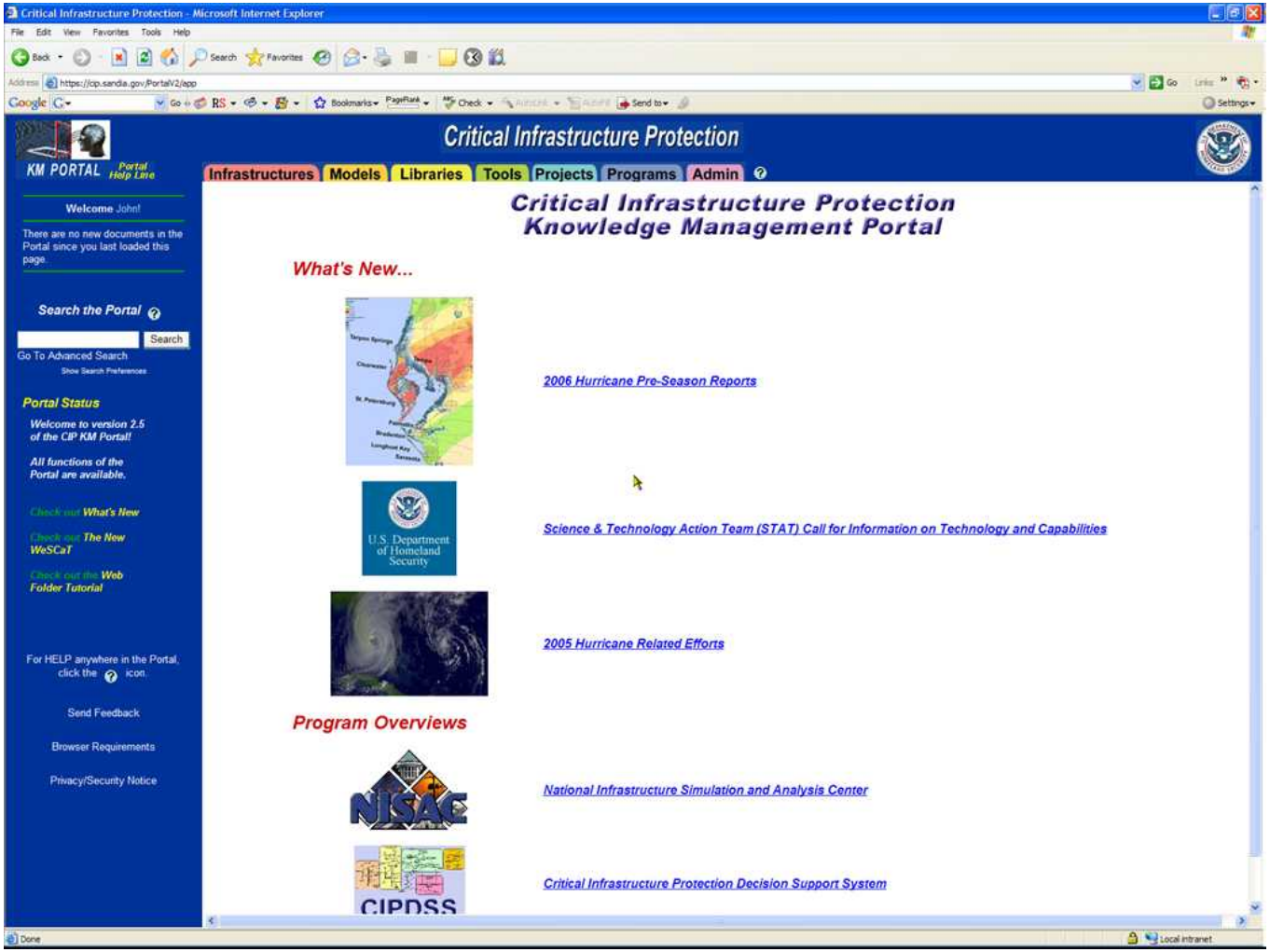
Los Alamos

Sandia National Laboratories

Why Was the Prototype Never Productionized?

- Funding downturn in the NISAC program
- Expense of semantically tagging almost 9,000 documents
  - With student help, over 10% were tagged in a few weeks
  - Semi-automatic approaches are being investigated
- Expense of re-implementing the portalCore interface using the Tapestry Web application framework
- Performance issues in determining document counts
  - Semantic Navigation examples on the Web generally assume that all information is available to everyone
  - Access group-based security schemes mean that document counts can potentially be different for every person, which is time-consuming to calculate on the fly

Synonym Expansion for Keyword Search

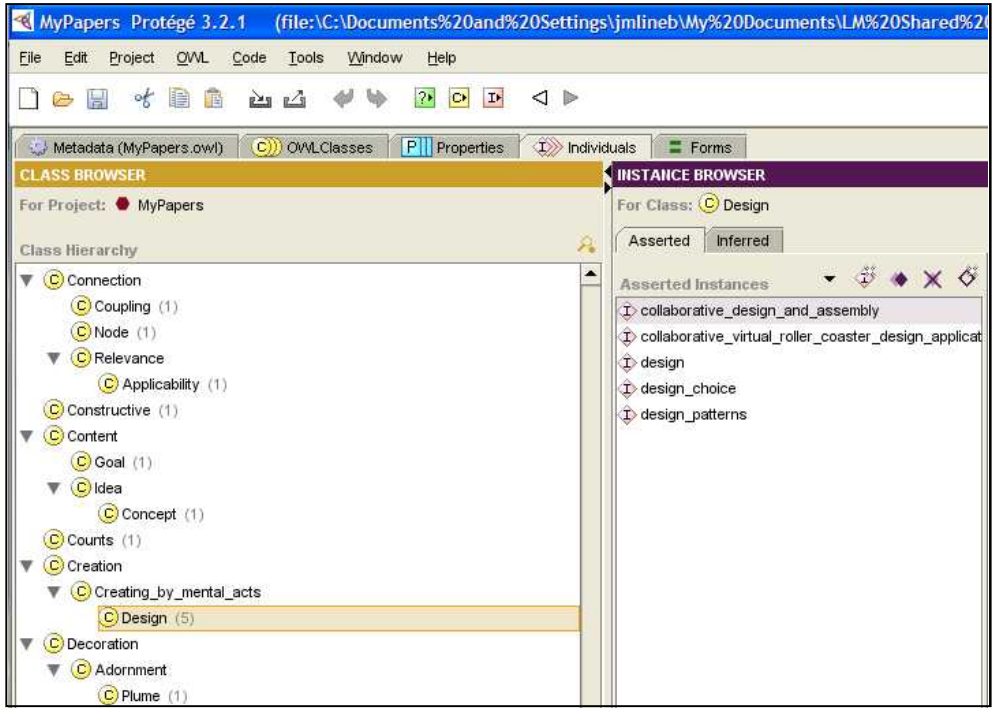


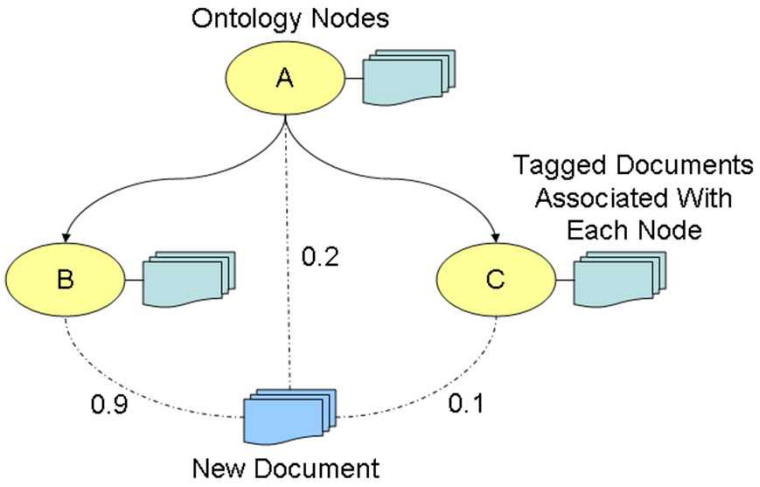
NISAC

Los Alamos

Sandia National Laboratories

Semantic Web Advanced Toolkit (SWAT)





Lockheed Martin-sponsored project to apply a statistical text analysis tool (STANLEY, for Sandia Text ANaLysis Extensible librarY) to the ontology development life-cycle

- Ontology learning from unstructured text
  - Classes (Entities)
  - Properties (*e.g.,* verbal relationships between entities, part-whole relations)
  - Upper-level ontology taken from WordNet
- Semi-automated semantic annotation
- Ontology evolution and maintenance

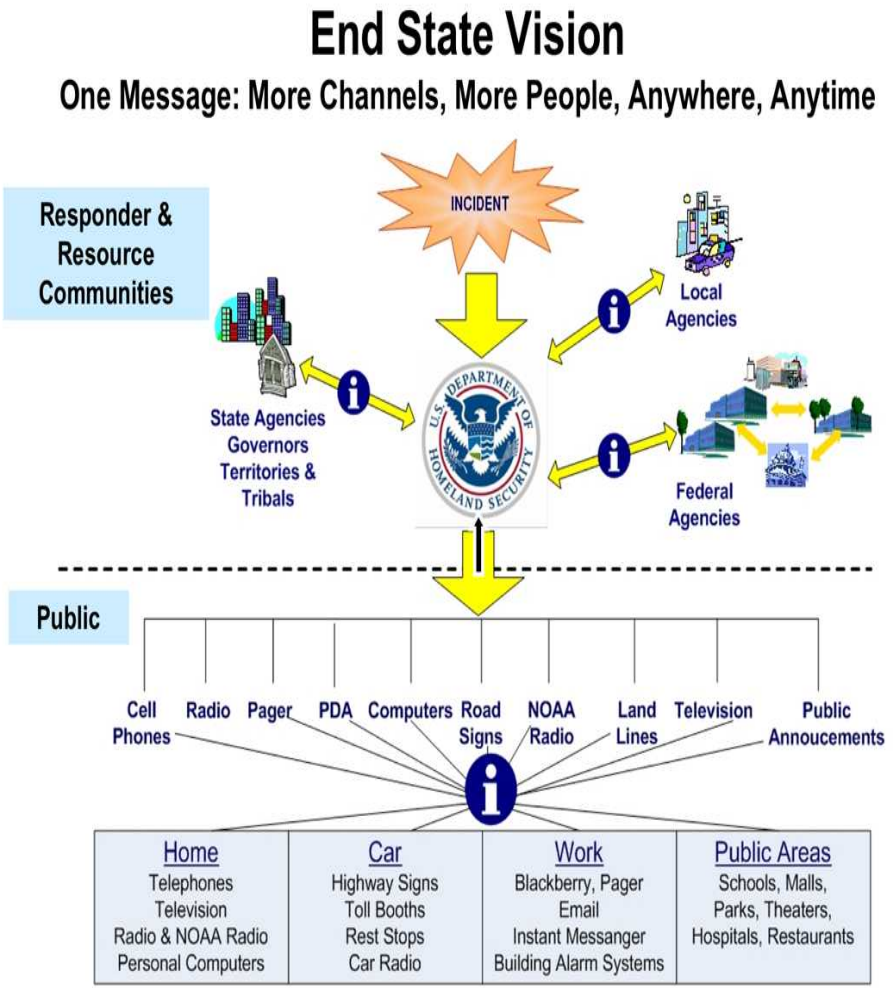
NISAC

Los Alamos

Sandia National Laboratories

FEMA IPAWS

- Integrated Public Alert and Warning System (IPAWS) for Federal Emergency Management Agency (FEMA)
- Ontology-based publish and subscribe mediation system for alert and warning messages
  - Messages published and subscribed to in terms of the ontology of each Community of Interest
  - Message routing done by mapping COI ontologies to a normative ontology

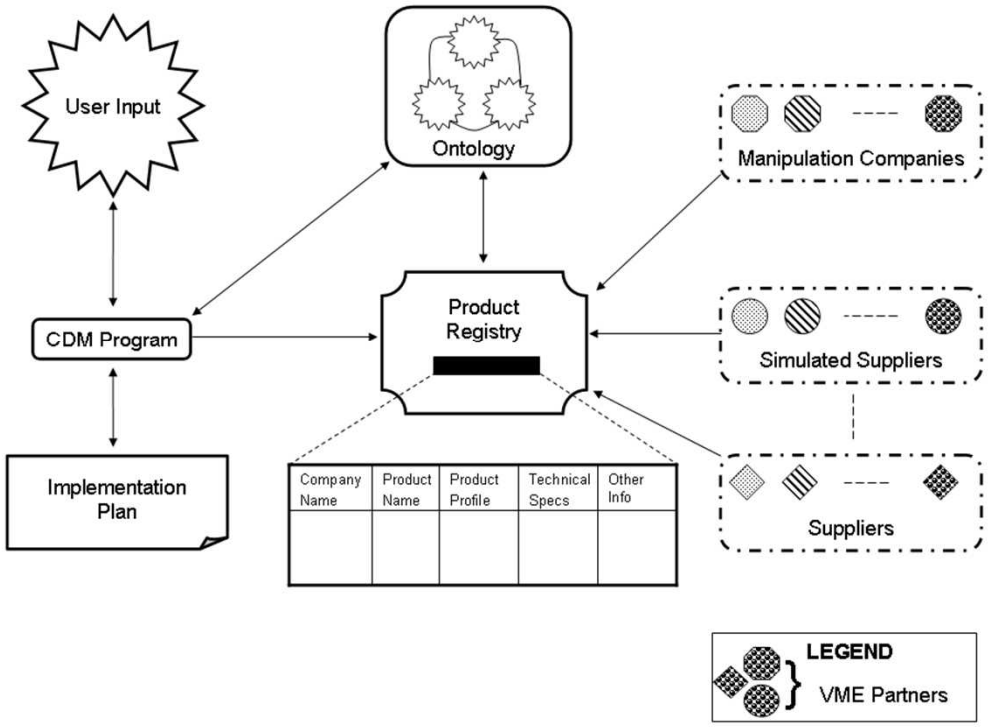


NISAC

Los Alamos

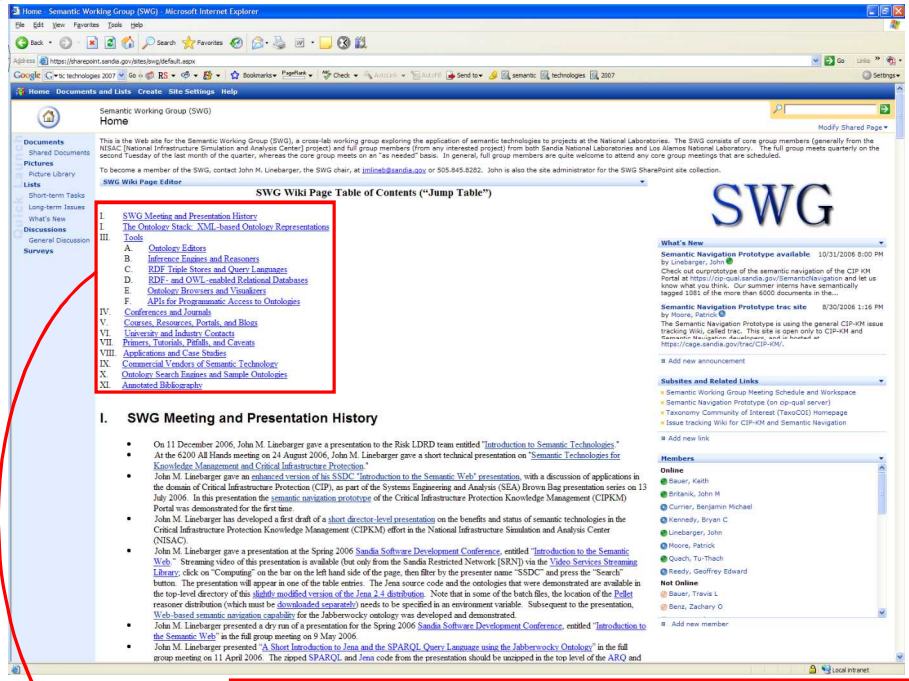
Sandia National Laboratories

Virtual Manufacturing Supply Chain



- Information integration of the virtual manufacturing supply chain in the nuclear weapons complex
- Small-scale virtual manufacturing enterprise (VME)
  - Ontology mapping
  - “Plug and play” semantic navigation

Organizational Recommendations



I. SWG Meeting and Presentation History
I. The Ontology Stack: XML-based Ontology Representations
III. Tools
A. Ontology Editors
B. Inference Engines and Reasoners
C. RDF Triple Stores and Query Languages
D. RDF- and OWL-enabled Relational Databases
E. Ontology Browsers and Visualizers
F. APIs for Programmatic Access to Ontologies
IV. Conferences and Journals
V. Courses, Resources, Portals, and Blogs
VI. University and Industry Contacts
VII. Primers, Tutorials, Pitfalls, and Caveats
VIII. Applications and Case Studies
IX. Commercial Vendors of Semantic Technology
X. Ontology Search Engines and Sample Ontologies
XI. Annotated Bibliography

- Attach yourself initially to an innovator or early adopter
- Begin with a "skunk works" team of 1-3 people working full-time
- Organize a semantics working group for knowledge sharing and set up a Wiki for it
- Don't underestimate the learning curve, or the difficulty of integrating into existing systems
- Train internally if possible, but may need to hire externally
- Prototype and productionize early and often
- Semantic metadata is expensive! Automate or support its creation.
- Pick problems people care about
- Must move beyond early adopters by demonstrating value to the wider organization to survive

NISAC

Los Alamos

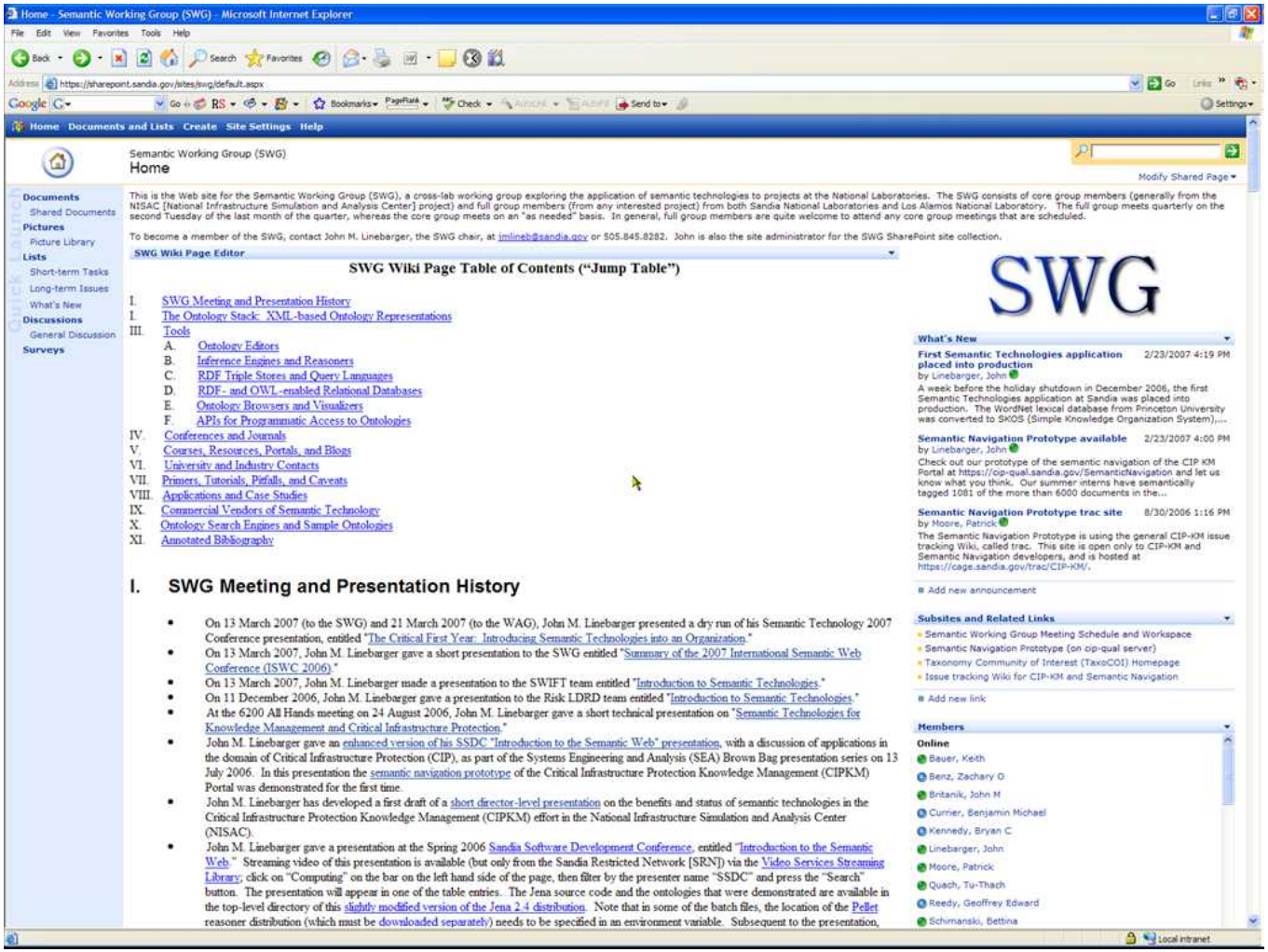
Sandia National Laboratories

Early Adopters

- They can find you
  - Working group Web site
  - Marketing materials
- You can find them
  - Project meetings
  - Political connections
- You can create them
  - Technology evangelism
  - On-line demonstrations of semantic technology from the annual Semantic Web Challenge
- But don't get too dependent upon them …

Image source: http://stylinonline.stores.yahoo.net

Semantic Working Group (SWG)



Los Alamos

Sandia National Laboratories

Learning Curve for Required Skill Sets

Semantic Navigation Prototype
OWL/RDFS/RDF
Protégé/Eclipse
Java Servlets
Jena API
Jena Rules
XSLT/XPath
Velocity Template Engine
Ant
Oracle

Synonym Expansion of Keywords
OWL/SKOS/ RDFS/RDF
Protégé/Eclipse
Java Servlets
Jena API
SPARQL
XSLT/XPath
WordNet API
Ant/Watir/log4j
Oracle

Semantic Web Advanced Toolkit
OWL/RDFS/RDF
Protégé/Visual Studio/Eclipse
C#/Java
Jena API
SPARQL
STANLEY API
WordNet API
NUnit
Pellet

Marketing Recommendations

- Practice (safe) promiscuous technology evangelism
  - Define, Explain, Differentiate, Demonstrate
- Such evangelism must be scalable
  - One-slide elevator speech
  - 15-minute management-level presentation
  - 2 hour technical presentation
- Articulate a vision for the rollout of semantic technologies in your company or organization
- Competitor technologies exist, so anticipate and address objections
- The Web site for your Working Group can also serve as a marketing tool
  - Demonstrations and Tools
  - Presentations and Publications
- Other important items
  - Glossy marketing materials
  - Motivating scenario
  - Pædogogical ontology and applications

NISAC Semantics Fact Sheet

Department of Homeland Security's Infrastructure Protection, Risk Management Division
National Infrastructure Simulation and Analysis Center (NISAC)

Semantic Technologies for Knowledge Synthesis

The National Infrastructure Simulation and Analysis Center (NISAC), a program under the Department of Homeland Security's (DHS) Preparedness Directorate, provides advanced modeling and simulation capabilities for the analysis of critical infrastructures, their interdependencies, vulnerabilities, and complexities. These capabilities help improve the robustness of our nation's critical infrastructures by aiding decision makers in the areas of preparedness, consequence and risk analysis, policy analysis, investment and mitigation planning, education and training, and near real-time assistance to crisis response organizations.

Sandia National Laboratories (SNL) and Los Alamos National Laboratory (LANL) are the prime contractors for NISAC, integrating the two laboratories' expertise in infrastructure disruption/vulnerability modeling and simulation under the direction of DHS's Infrastructure Protection/Risk Management Division.

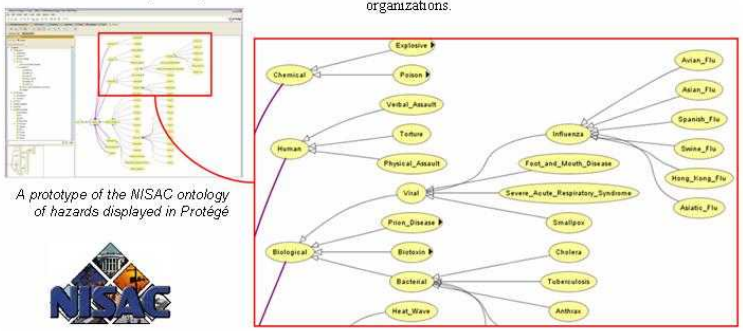
Knowledge Synthesis

NISAC analysts and modelers require vast amounts of data and information for their critical infrastructure interdependency work. The knowledge synthesis effort provides an access-controlled, organized, and searchable view on the data for rapid retrieval and update through a suite of Web-based tools and programmatic interfaces. Four semantic technology projects are underway to help improve information access:

- Ontology Development
- Semantic Navigation
- Keyword Expansion (Synonyms)
- Automatic Semantic Metadata Determination

Ontology Development

An ontology is a map of the key terms in a domain of knowledge and the relationships among the terms. It forms both a vocabulary and a cognitive framework for reasoning about the domain. By structuring domain knowledge and making relationships both within and among subject areas explicit, NISAC analysts are able to categorize and access information more efficiently. Additionally, the ontology is flexible enough to be adjusted as the understanding of the domain evolves. High level structural components of the current ontology for the critical infrastructure protection domain include hazards, critical assets, geographical location, consequences, and organizations.



A prototype of the NISAC ontology of hazards displayed in Protégé

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Semantic Navigation

The NISAC Critical Infrastructure Protection Knowledge Management (CIP KM) Portal is a one-stop shop where NISAC analysts, modelers, and administrators can contribute and share information. Semantic navigation allows the Knowledge Management Portal to be searched in terms of concepts and relationships, rather than by keyword only. This approach overcomes the semantic ambiguity and unrecognized synonym issues associated with keyword search. As a result, analysts can navigate the portal at the "speed of click" using categories drawn from the domain of critical infrastructure protection.

Web-based Semantic Navigation of the CIP KM Portal

Keyword Expansion

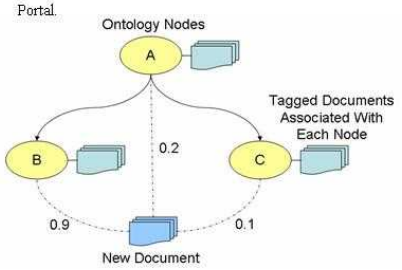
A step along the way to full synonym resolution is the expansion of keywords used to search the CIP KM Portal. Each keyword is expanded to include synonyms related to the concept being searched, in priority order. For example, a search on the term "bird flu" brings back documents containing that phrase and also those containing "avian influenza" and "H5N1." The synonyms are drawn from those commonly used by critical infrastructure protection analysts, as well as a more general set of synonyms taken from a public domain thesaurus.

Automatic Semantic Metadata Determination

As new documents are added to the CIP KM Portal, they must be tagged under metadata categories defined within the ontology in order to be available for semantic navigation. This is a tedious and error-prone process when carried out manually. A system is being developed to automatically recommend the appropriate metadata categories using a text analysis tool. However, the person submitting the information will review the recommendations and make the final decision. The benefits of automatic semantic metadata determination include:

- Efficient addition of new documents, which minimizes human involvement in determining semantic metadata
- Potentially better quality and consistency of semantic metadata

An extension of this application is being designed to target information on the Internet, making a tremendous amount of other data resources available to users of the CIP KM Portal.



A new document comes in and shows the highest correlation (0.9) with node B of the ontology. The system recommends that the document be semantically tagged using the semantic metadata of that node.

Contacts:

Jon MacLaren
DHS-IP
(703) 235-5489; email:
jon.m.maclaren@dhs.gov

Theresa Brown
Sandia National Laboratories
(505) 844-5247; email:
tjbrown@sandia.gov

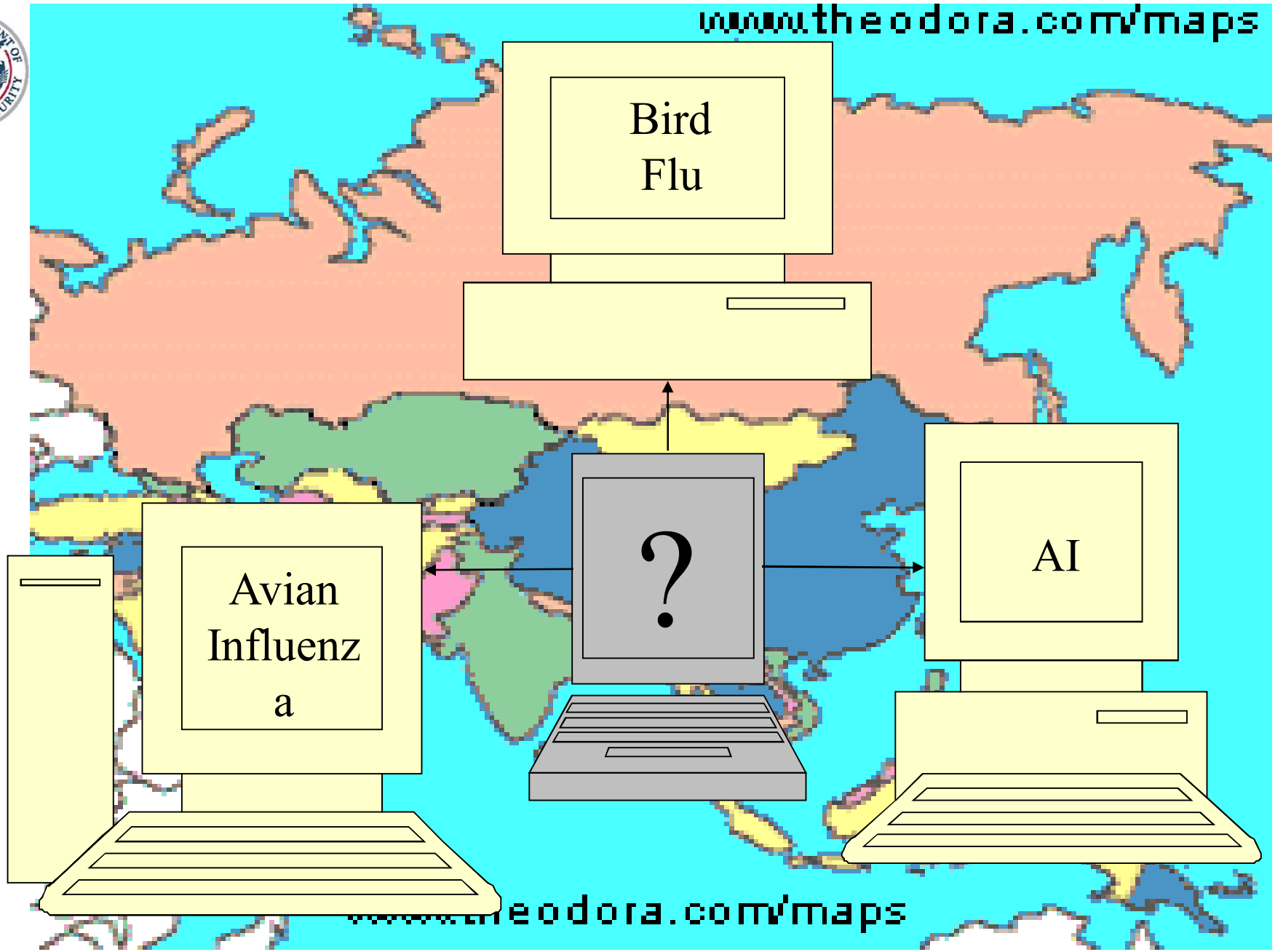
NISAC

Los Alamos

Sandia National Laboratories

Motivating Scenario

- Imagine a team of employees collaborating on-line on a time-critical analysis of avian flu. One employee posts a document under 'Avian Influenza,' while another marks theirs as 'Bird Flu.' A third, pressed for time, abbreviates the topic as 'AI.' Each person uses his or her own filing system and organizational taxonomy, and when a keyword search is performed, the computer has no idea that 'Avian Influenza' is 'Bird Flu' is 'AI.'
- With deadlines looming, no one can find the others' documents, at least by keyword, leading to stress, delays, and unnecessary headaches. This is an example of the Semantic Problem.



NISAC

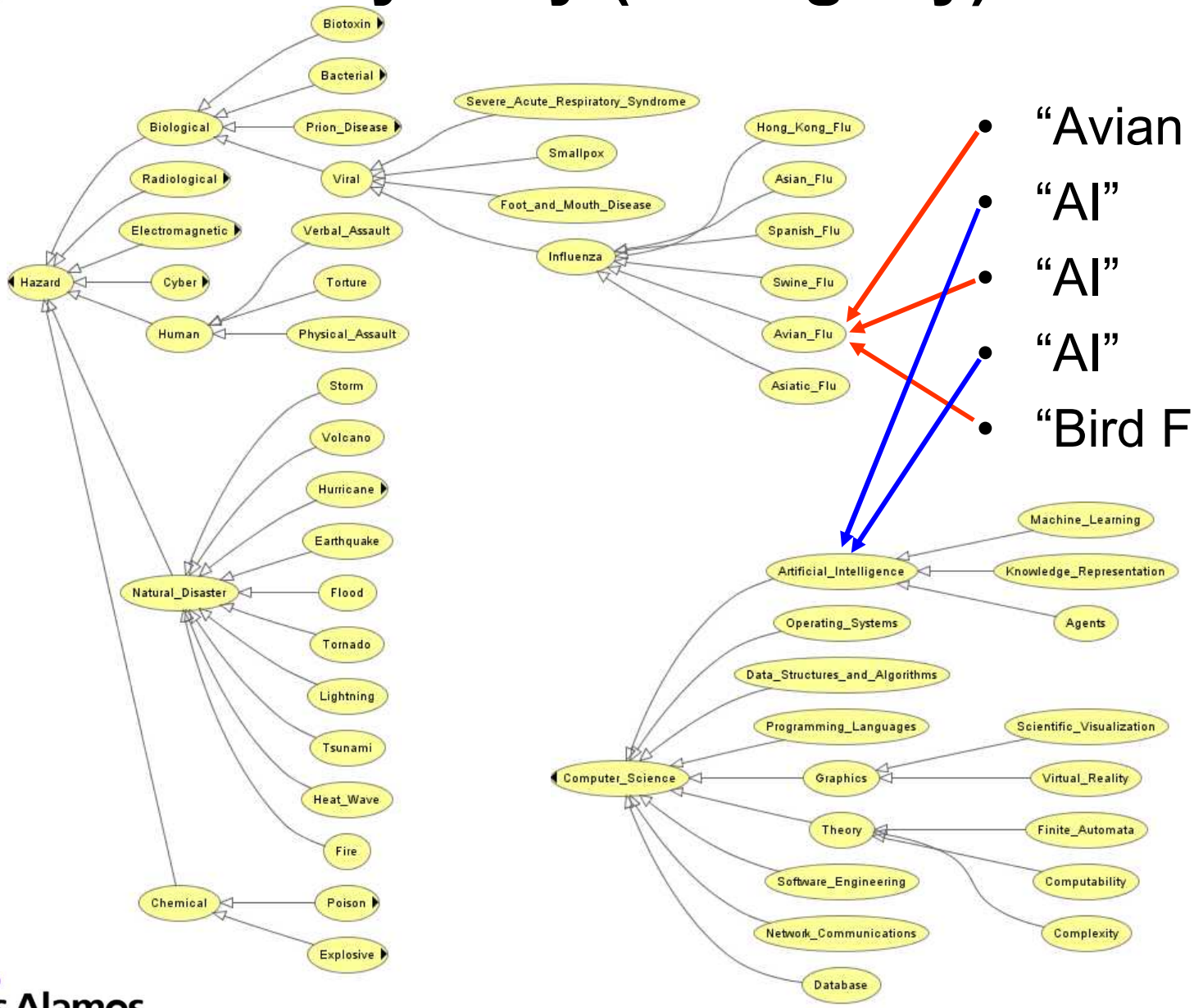
Los Alamos

Sandia National Laboratories

Synonym Resolution

- "Avian Flu"
- "AI"
- "Bird Flu"

Polysemy (Ambiguity) Resolution

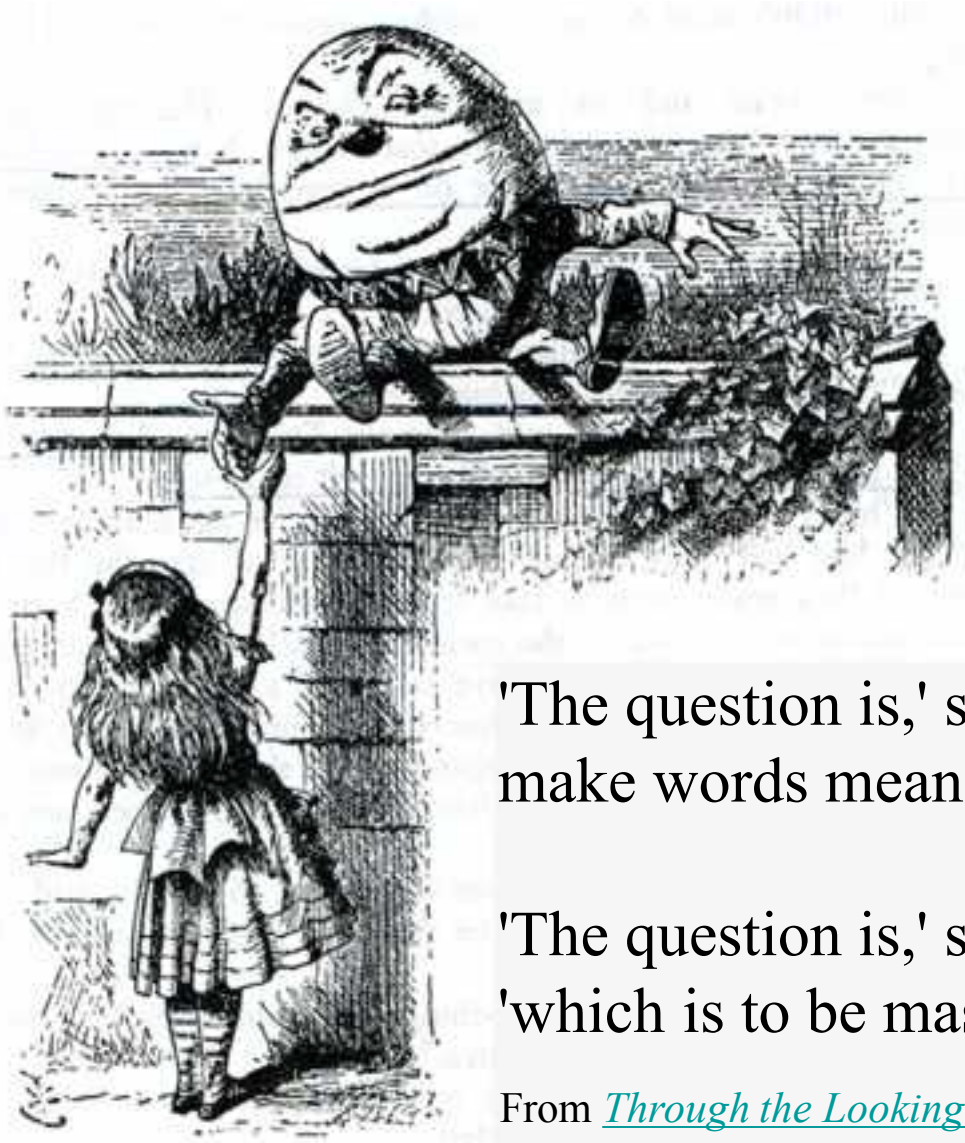


NISAC

Los Alamos

Sandia National Laboratories

The Semantic Problem



'When **I** use a word,' Humpty Dumpty said, in a rather scornful tone,' it means just what I choose it to mean, neither more nor less.'

'The question is,' said Alice, 'whether you *can* make words mean so many different things.'

'The question is,' said Humpty Dumpty, 'which is to be master - that's all.'

From *Through the Looking Glass* by Lewis Carroll

NISAC

Los Alamos

Sandia National Laboratories

Jabberwocky

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

"Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!"

He took his vorpal blade in hand;
Long time the manxome foe he sought --
So rested he by the Tumtum tree,
And stood a while in thought.

And, as in uffish thought he stood,
The Jabberwock, with eyes of flame,
Came whiffling through the tulgey wood,
And burbled as it came!

One, two! One, two! And through and through
The vorpal blade went snicker-snack!
He left it dead, and with its head
He went galumphing back.

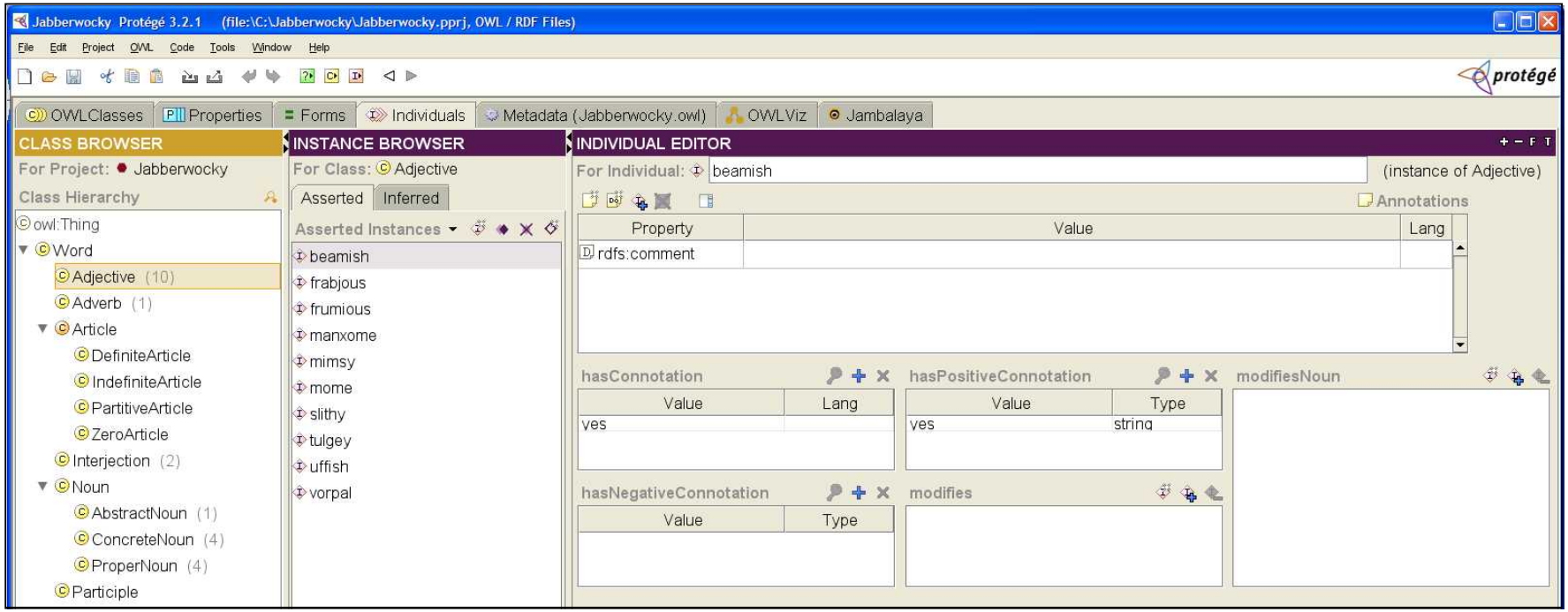
"And hast thou slain the Jabberwock?
Come to my arms, my beamish boy!
O frabjous day! Callooh, Callay!"
He chortled in his joy.

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

Jabberwocky Ontology

- Ontology of parts of speech
  - Nouns, Verbs, Adverbs, Adjectives, Interjections
  - Class instances were used for denotations, properties were used for connotations
    - Positive
    - Negative
- Notice that Jabberwocky is actually a dialect of English; another dialect of English is Scots
- A SPARQL query was developed to randomly combine Jabberwocky and Scots words into sentences with positive connotations and sentences with negative connotations
- Semantic navigation capability was demonstrated using the portalCore framework from the SWAD-E project

Jabberwocky Ontology



Browse

The data in this data source is organized according to the characteristics or facets which are shown below. The numbers in parentheses show how many results are in that facet. To start browsing, pick an option from one of the facets.

Click ⊞ to see full lists of options for each facet.

Word Type

Adjective (26) | Adverb (1) | Interjection (2) | Noun (26) | Verb (8)

Word Dialect

Jabberwocky (28) | Scots (34)

Word Name

A* (1) | B* (9) | C* (5) | D* (2) | E* (1) | F* (3) | G* (9) | H* (2) | J* (4) | K* (4) | L* (3) | M* (4) | N* (1) | O* (1) | R* (2) | S* (4) | T* (3) | U* (1) | V* (1) | W* (2)

Language Dialect

English (11) | French (7) | Spanish (5)

Los Alamos

Sandia National Laboratories

Ontology Development Issues

- Ontology development is not the same as software development
  - Knowledge engineering is not software engineering
  - Special skills and training are needed
- Take the time to learn the formal Description Logic underpinnings of OWL; there are good courses and tutorials on the Web.
- Reuse instead of reinventing
  - *e.g.*, FOAF, BibTeX, Dublin Core
- Consider using appropriate upper ontologies
  - *e.g.*, Cyc, SUMO, WordNet
- Beware the open world assumption!
  - Quantified restrictions are tricky
  - For example, `hasNegativeConnotation has "yes"` and `hasPositiveConnotation exactly 0` did not classify Impolite Words in the Jabberwocky ontology

NISAC

Los Alamos

Sandia National Laboratories

How to Learn Ontology Development

- Formal degree in Knowledge Representation or Knowledge Engineering
- Short courses, tutorials, and seminars on ontology development (which may be vendor- or tool-specific)
- Self-teaching resources
  - "How to Build an Ontology" video from University of Buffalo
  - Tutorials and papers from Manchester and Maryland
    - "Practical Guide" and "Common Errors and Patterns"
    - "Debugging Owl Ontologies" Web page and paper(s)
  - "Learn from the Masters" by studying existing ontologies
    - Presentation materials from ISWC 2006 tutorial
    - Videos of the ISWC 2006 tutorial
  - *Ontological Engineering* book by Gómez-Pérez *et al.*
- Tip: Develop your ontology with a reasoner, such that the structure of your ontology is not statically determined but instead is inferred by the reasoner

Tools and Support Issues

- Start with free or open source tools
  - Protégé has newsgroups with RSS feeds
  - Jena has an extremely responsive Yahoo support group with an RSS feed
  - Sesame has support forums and mailing lists
  - Pellet has mailing list
  - Saxon parser for XSLT/XPath has a mailing list and forum
- It's still a small world after all
  - Seek out contacts with movers and shakers
  - "Strip-mine" conferences and workshops
    - International Semantic Web Conference (ISWC)
    - Semantic Technologies Conference
  - "Cold call" email requests for information are generally received favorably, especially if you have information to offer in return

Resources

- The May 2001 Scientific American article that started it all: "The Semantic Web," by Tim Berners-Lee, James Hendler, and Ora Lassila
- The 2006 update to the above article, in IEEE Intelligent Systems: "The Semantic Web Revisited," by Nigel Shadbolt, Wendy Hall, and Tim Berners-Lee
- *A Semantic Web Primer* by Grigoris Antoniou and Frank van Harmelen
- "A Practical Guide to Building OWL Ontologies" (a.k.a. "Manchester Pizza Tutorial") for OWL and ontology development
- Swoogle to search for existing ontologies

Acknowledgements

- The National Infrastructure Simulation and Analysis Center (NISAC) is a program under the Department of Homeland Security's (DHS) Preparedness Directorate. Sandia National Laboratories (SNL) and Los Alamos National Laboratory (LANL) are the prime contractors for NISAC under the programmatic direction of DHS's Infrastructure Protection/Risk Management Division.
- Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

NISAC

Los Alamos

Sandia National Laboratories

NISAC

Question and Answer

Los Alamos

Sandia National Laboratories