Ontology-Based Applications in the Age of the Semantic Web

Prof Enrico Motta, *PhD*

Knowledge Media Institute
The Open University
Milton Keynes, UK

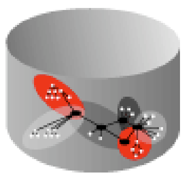
The Rise of Semantics

ORACLE
TECHNOLOGY NETWORK

PRODUCTS
Database
Middleware
Developer Tools
Enterprise Management
Applications Technology
Extensions and Plugins
Products A-Z

TECHNOLOGIES
BI & Data Warehousing
Java
Linux

Getting Started | Downloads | Documentation | Forums



Semantic Technologies Center

Semantic Technologies are designed to extend the capabilities of infc
enterprise systems to be networked in meaningful ways. The adoptio
Consortium (W3C) standards like XML, RDF (Resource Description F
Ontology Language) serve as foundation technologies to advancing t
technologies.

Oracle Spatial 10*g* introduces the industry's first open, scalab
graph data model, RDF triples are persisted, indexed and qu
10*g* RDF database ensures that application developers bene
secure semantic applications. Application areas include:

Gartner

Finding and Exploiting Value in Semantic Technologies on the Web

9 May 2007

David W. Cearley   Whit Andrews   Nicholas Gall

)atabase to design and de

garlik™

Home
About Garlik
DataPatrol
Advice
Sign up

powerfulstuff   sign

Revolutionary online protection

DataPatrol is a new monthly
monitoring service that finds, tracks

IBM®

Home | Products | Services & industry solutions | Support & downloads | My IBM

developerWorks

In this article:

- Technologies that make up the Semantic Web

developerWorks > Web development | Open source | XML >

The future of the Web is Semantic

Ontologies form the backbone of a whole new way to understand online data

Level: Introductory

NeOn   The Open University

Increasing Semantic Content

DBpedia.org

Querying Wikipedia like a Database.

Enrico Motta
Professor of Knowledge Technologies [info] [homepage] [e

Liza Mu
Visiting Researcher [info] [email] [RDF/XML]

```
<rdf:Description rdf:about="">
     <mediapro:People>
        <rdf:Bag>
          <rdf:li>Jim Hendler</rdf:li>
          <rdf:li>Enrico</rdf:li>
        </rdf:Bag>
     </mediapro:People>
```

AKT Triplestore Browser

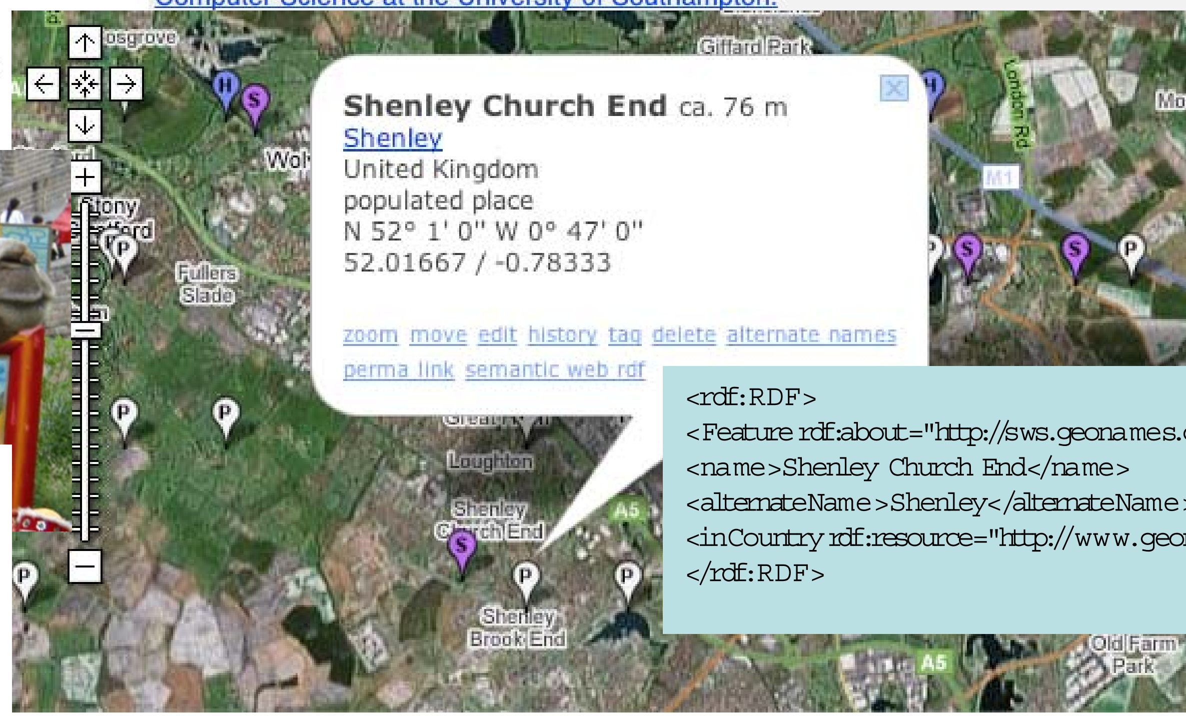
about browse manage query demo

Southampton ECS People

file:/usr/local/share/akt/Southampton/southampton-people.rdf

creator AKT Project
Nick Gibbins

description This ontology contains information about the members of the Department of Electronics and Computer Science at the University of Southampton.




```
<rdf:RDF>
<Feature rdf:about="http://sws.geonames.org/26380
<name>Shenley Church End</name>
<alternateName>Shenley</alternateName>
<inCountry rdf:resource="http://www.geonames.org
</rdf:RDF>
```

The Open University

Key Propositions

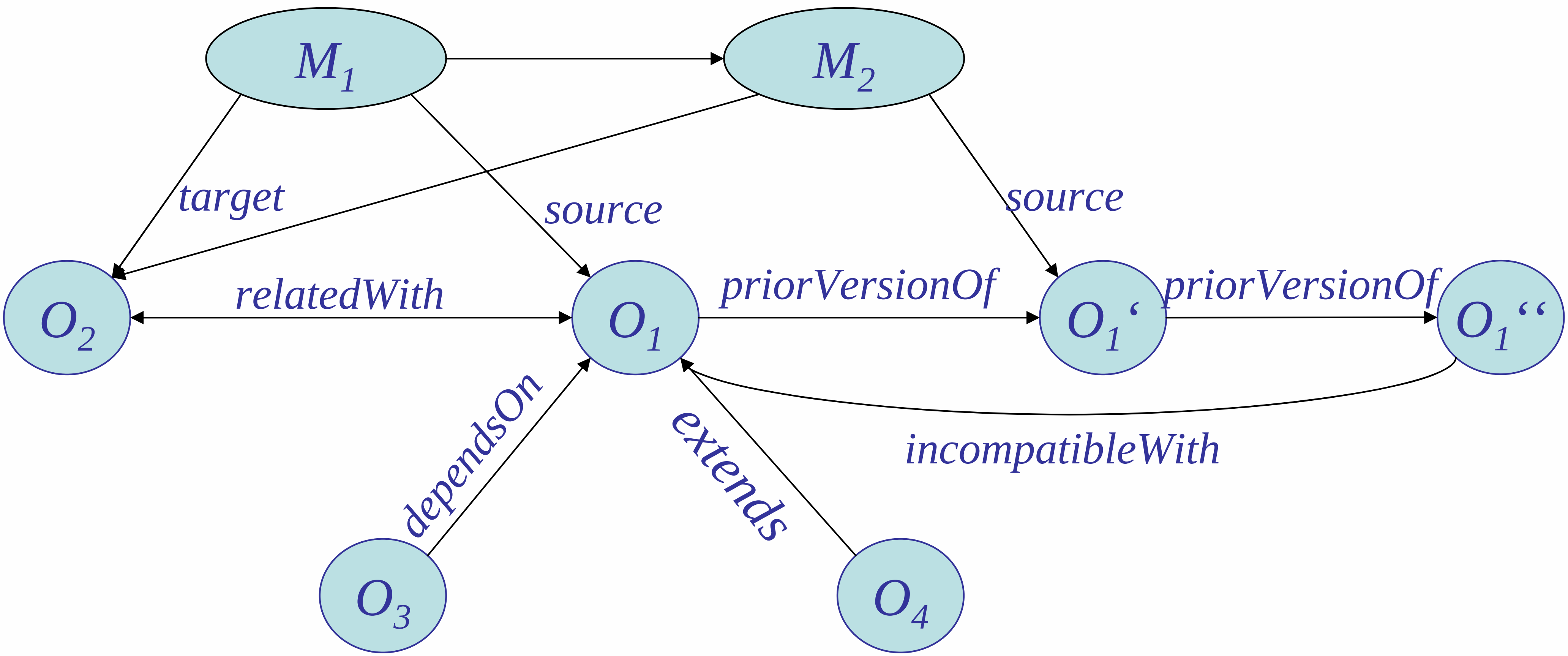
- The SW is less and less an aspiration and more and more a reality
- This emerging large scale semantics opens up new scenarios and introduces a number of implications for:
  – the practice of ontology engineering
  – the kind of functionalities that ontology engineering tools ought to support
  – the kind of ontology-based applications we can now develop
- In addition, it may also provide a solution to one of the holy grails of AI research: the availability of large-scale background knowledge to enable intelligent behaviour

Large Scale Semantics and Ontology Engineering

Ontology Engineering in the Age of the Semantic Web

- The availability of large scale semantics (millions of docs and tens of thousands of ontologies) opens up the following opportunities
  - to make cost-effective the develoment of large scale semantic applications out of reusable resources
  - to move away from monolithic ontologies and characterise ontology engineering as the process of constructing and managing ***networked ontologies***
- The goal of the NeOn project is precisely to provide a methodology and a novel infrastructure for ontology engineering in line with this vision of the next generation of ontology-based applications

Networked Ontologies



First Year Outputs: Some Highlights

- Meta model and initial methods for reasoning with networked ontologies
- A formal, ontological framework for characterizing collaborative ontology design workflows
- Formalization of context and initial methods and software for generating mappings which contextualise ontologies
- New methods for ontology alignment, selection and modularization
- **A task-centric user study highlighting limitations of current tools in tackling typical NeOn development scenarios**
- Initial modelling components for NeOn methodology
- NeOn Architecture design and initial infrastructure components
- Initial Version of the NeOn Toolkit
- Analysis and design of NeOn testbeds

http://www.neon-project.org/

Outline of the study

- 2 ontology engineering tools
  - TopBraid, Protégé
- 3 ontologies
  - Copyright (85 C; 49 P; 128 Re)
  - AKT Support (14 C; 15 P; 0 Re)
  - AKT Portal (162 C; 120 P; 130 Re)
- 28 participants
  - Mixed w.r.t. expertise with ontologies and tools
  - Actually most users had designed ontologies in the past, but usually not in OWL
- 3 tasks
  - Task 1: Simple class/subclass relationship across ontologies
  - Task 2: Import two ontologies and change axioms
  - Task 3: Import concepts and redefine them

Attitudes towards NeOn functionalities

Question (existing feature or 'proposed fix')	Avg. marks	–1	0	+1	Total
Existing support for ontology re-use	–0.097 (not very good / reasonable)	26%	58%	16%	31
Support for partial re-use of ontologies	–0.739 (not very good)	62%	14%	4%	29
flag chunks of ontologies or concept worked with	+0.674 (*would be useful)*	20%	24%	56%	25
hide selected (irrelevant?) parts of ontologies	+0.465 (*would be reasonable / useful)*	25%	38%	38%	24
Existing support for mappings and contextual boundaries	–0.065 (not very good / reasonable)	19%	68%	13%	31
Management and assistance with any mappings	–0.480 (not very good / reasonable)	48%	52%	0%	26
propose mappings & ensure their consistency	+0.433 (*would be reasonable/useful)*	3%	50%	47%	30
using trial queries to see consequences of mappings	+0.045 (*would be reasonable)*	9%	77%	14%	23
Existing support for versioning, alternatives	–0.200 (not very good)	50%	20%	30%	11
Existing visualizing capabilities & their adaptation	–0.536 (not very good)	57%	39%	4%	28
propagate changes between alternative versions	+0.519 (*would be reasonable / useful)*	7%	33%	60%	28
compare/visualize different interpretations/versions	+0.700 (*would be useful)*	6%	17%	77%	30
performing operations in graphical/textual mode	+0.414 (*would be reasonable / useful)*	7%	45%	48%	29
visualize on the level of ontologies (not just concepts)	+0.357 (*would be reasonable / useful)*	7%	50%	43%	28

Implications for ontology engineering infrastructure

- Empirical findings confirm intuition that existing tools need new functionalities to support the NeOn vision
- This is potentially a critical issue as the tension between what is feasible in principle and what is supported by the current infrastructure may generate a “software crisis”.
- Problems are clearly harder for less expert users, which actually provide the key industrial target audience

Implications for Ontology Engineering Practice

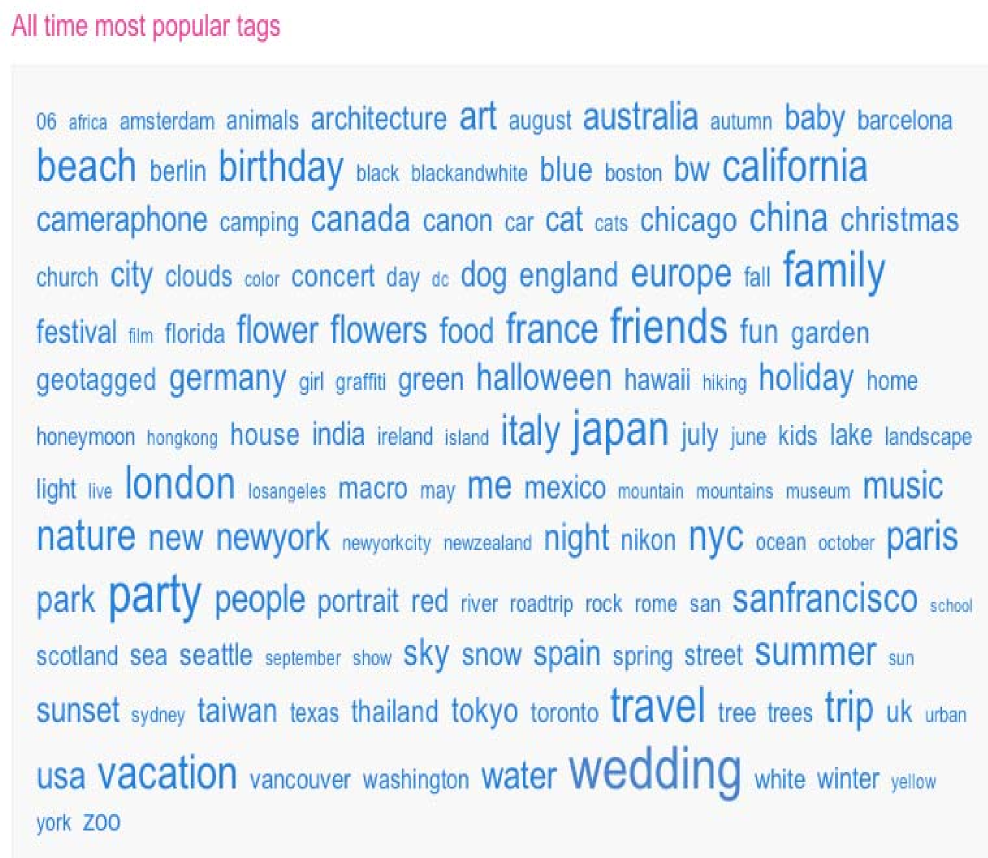
- Reuse rather than ad hoc design of ontology elements
- Potential for making the Ont. Dvpt. process more robust
  - Cfr. similar paradigm shift for KBS thanks to work on Problem Solving Methods (1985 - onwards)
- The NeOn vision nicely complements ongoing work on design patterns for ontology engineering
  - Meta-level nature of design patterns vs.
    object-level nature of direct reuse of definitions
  - NeOn methodology is indeed based on work on design patterns
- Ontology engineering generates new kinds of outputs
  - Networked ontologies
    - when process creates connections between distributed pre-existing ontologies
  - Faceted ontologies
    - when process consists of creating a new ontology out of massively distributed ‘ontology snippets’

Faceted Ontologies

Example: Integrating SW and Web2.0

Features of Web2.0 sites

- Tagging as opposed to rigid classification
- Dynamic vocabulary does not require much annotation effort and evolves easily
- Shared vocabulary emerge over time
  - certain tags become particularly popular



Limitations of tagging

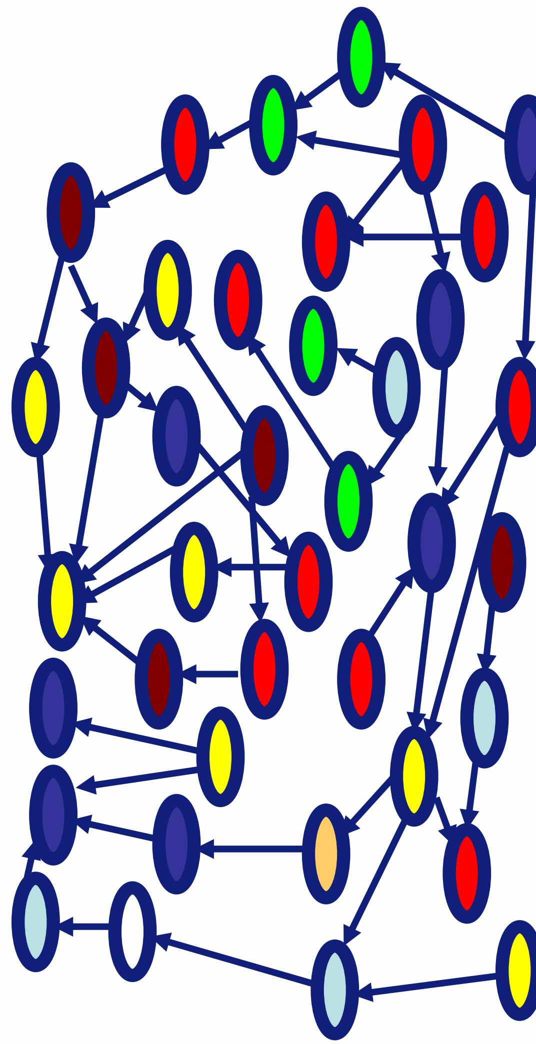
- Different granularity of tagging
  - rome vs colosseum vs roman monument
  - Flower vs tulip
  - Etc..
- Multilinguality
- Spelling errors, different terminology, plural vs singular, etc...

- This has a number of negative implications for the effective use of tagged resources
  - e.g., Search exhibits very poor recall

Giving meaning to tags

All time most popular tags

06 africa amsterdam animals architecture art august australia autumn baby barcelona beach berlin birthday black blackandwhite blue boston bw california cameraphone camping canada canon car cat cats chicago china christmas church city clouds color concert day dc dog england europe fall family festival film florida flower flowers food france friends fun garden geotagged germany girl graffiti green halloween hawaii hiking holiday home honeymoon hongkong house india ireland island italy japan july june kids lake landscape light live london losangeles macro may me mexico mountain mountains museum music nature new newyork newyorkcity newzealand night nikon nyc ocean october paris park party people portrait red river roadtrip rock rome san sanfrancisco school scotland sea seattle september show sky snow spain spring street summer sun sunset sydney taiwan texas thailand tokyo toronto travel tree trees trip uk urban usa vacation vancouver washington water wedding white winter yellow york zoo



What does it mean to add semantics to tags?

1. Mapping a tag to a SW element

"japan"

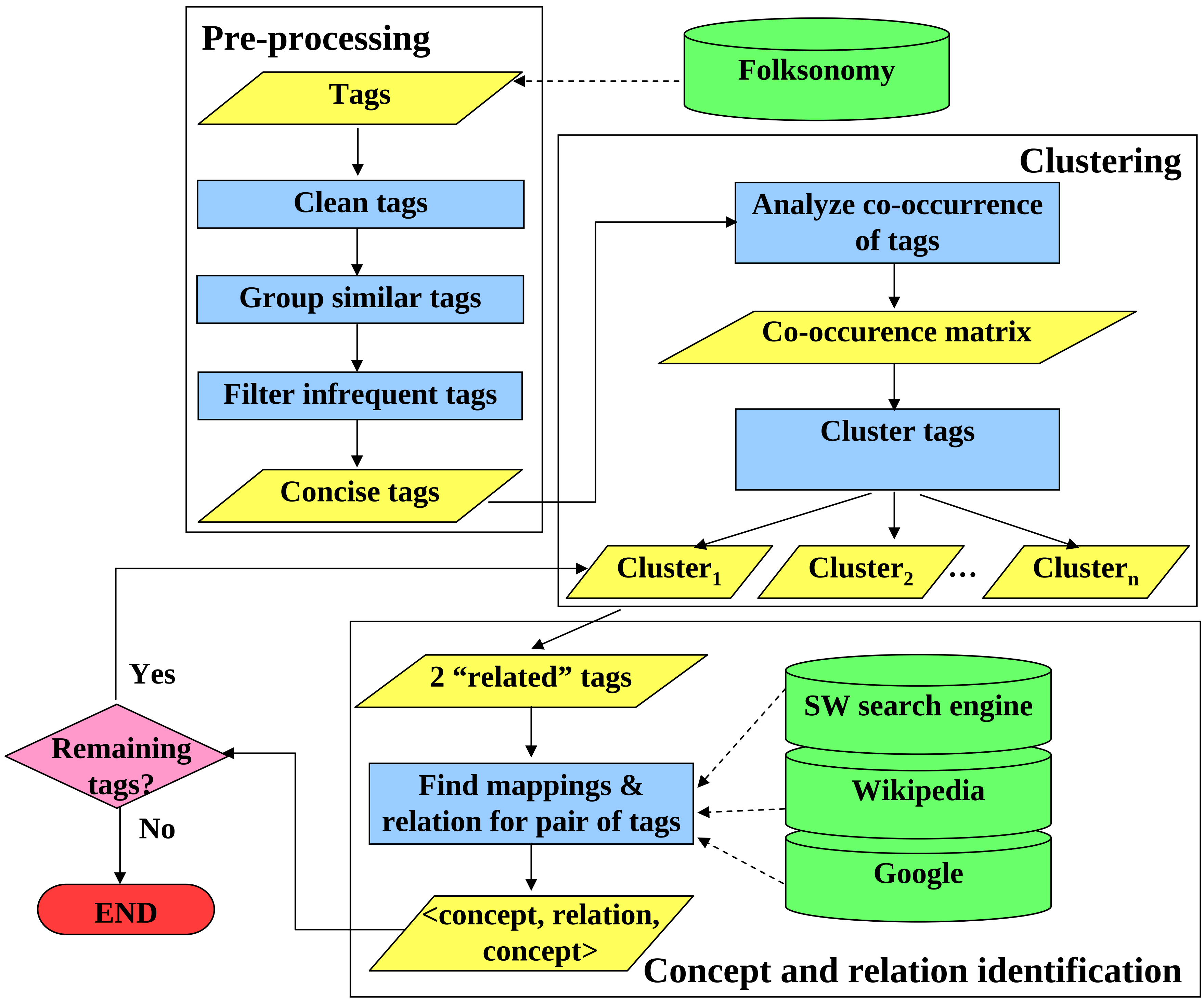
<akt:Country Japan>

2. Linking two "SW tags" using semantic relations

{japan, asia} → <japan subRegionOf asia>

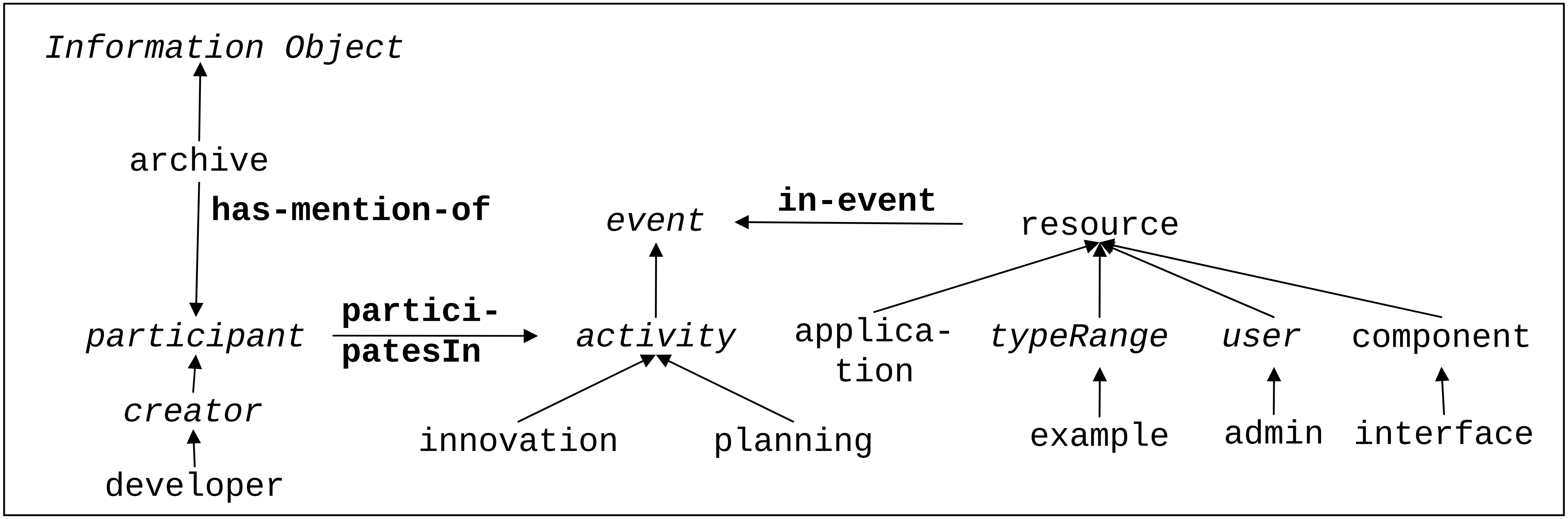
Applications of the approach

- To improve recall in keyword search
- To support annotation by dynamically suggesting relevant tags or visualizing the structure of relevant tags
- To enable formal queries over a space of tags
  - Hence, going beyond keyword search
- To support new forms of intelligent navigation
  - i.e., using the 'semantic layer' to support navigation


Pre-processing
Tags
Clean tags
Group similar tags
Filter infrequent tags
Concise tags
Folksonomy
Clustering
Analyze co-occurrence of tags
Co-occurence matrix
Cluster tags
Cluster$_1$
Cluster$_2$
…
Cluster$_n$
Yes
Remaining tags?
No
END
2 “related” tags
Find mappings & relation for pair of tags
SW search engine
Wikipedia
Google
<concept, relation, concept>
Concept and relation identification

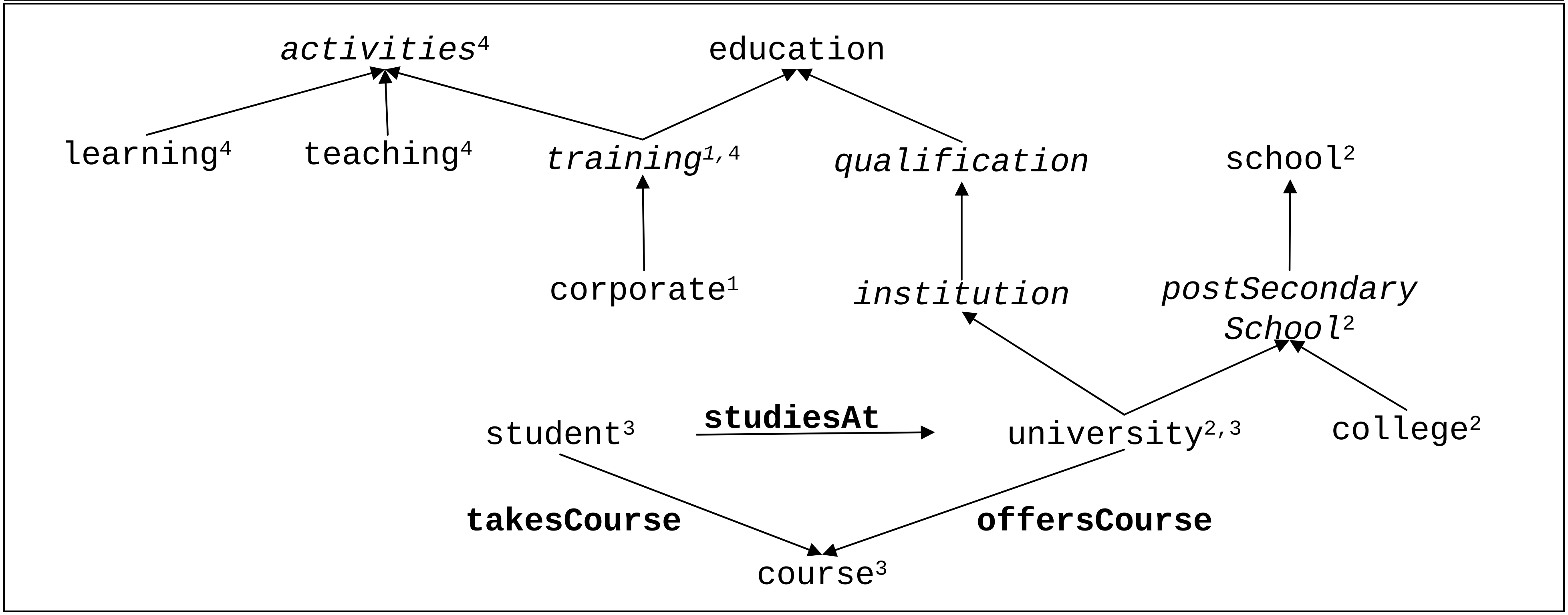
Examples

Cluster_1: {admin application archive collection component control developer dom example form innovation interface layout planning program repository resource sourcecode}



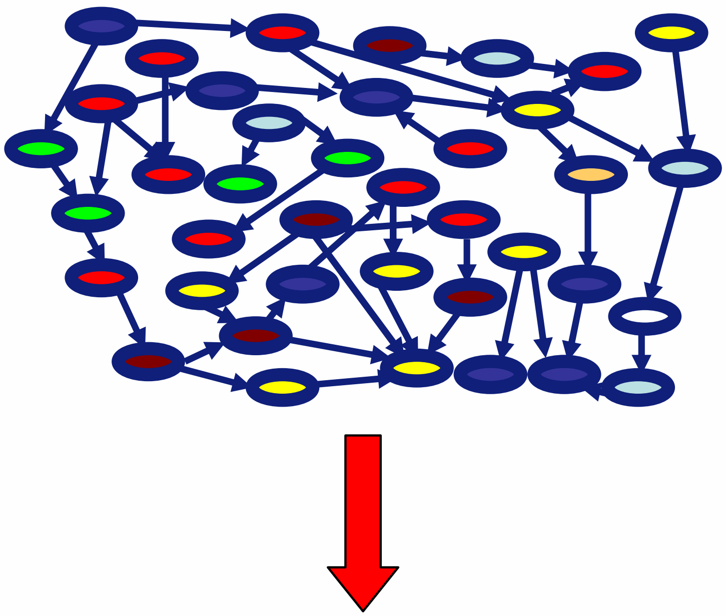
Examples

Cluster_2: {college commerce corporate course education high instructing learn learning lms school student}



[1]http://gate.ac.uk/projects/htechsight/Employment.daml.
[2]http://reliant.teknowledge.com/DAML/Mid-level-ontology.daml.
[3]http://www.mondeca.com/owl/moses/ita.owl.
[4]http://www.cs.utexas.edu/users/mfkb/RKF/tree/CLib-core-office.owl.

Faceted Ontology



- Ontology creation and maintenance is automated
- Ontology evolution is driven by task features and by user changes
- Large scale integration of ontology elements from massively distributed online ontologies
- Very different from traditional top-down-designed ontologies

Second Generation Semantic Web Applications

- The example given provides an example of a new generation of SW applications, with the following features:
  - Dynamic use of online knowledge
  - SW is used as a large scale repository providing background knowledge to an intelligent problem solver
  - No single ontology driving data integration

- The new class of systems enabled by the SW is fundamentally different in many respects both from traditional KBS and even from early SW applications

- The difference between 1st and 2nd generation SW applications can be seen as that between “corporate semantic webs” and “open semantic web”

CS AKTive Space

Take a tour

About this page ◉ research area/region ○ region/research area

Research area

information interfaces and presentation
information systems applications
information storage and retrieval
database management
general
Computing Methodologies
document and text processing
simulation and modeling
pattern recognition
image processing and computer vision
computer graphics
artificial intelligence
symbolic and algebraic manipulation
general
Computer Applications
computers in other systems
computer-aided engineering
arts and humanities
social and behavioral sciences
life and medical sciences

Radial: 100 miles

Map: uk-political

Researcher

Top ◉ 5 ○ 10 ○ 20 ○ unlimited

Order by ☑ Grant total ☐ RAE result

NR Shadbolt
LA Carr
DC De Roure
NR Jennings
L Moreau

Overview: NR Shadbolt

browse

Name NR Shadbolt

Institution Intelligence, Agents and Multimedia, University of Southampton

Email nrs@ecs.soton.ac.uk

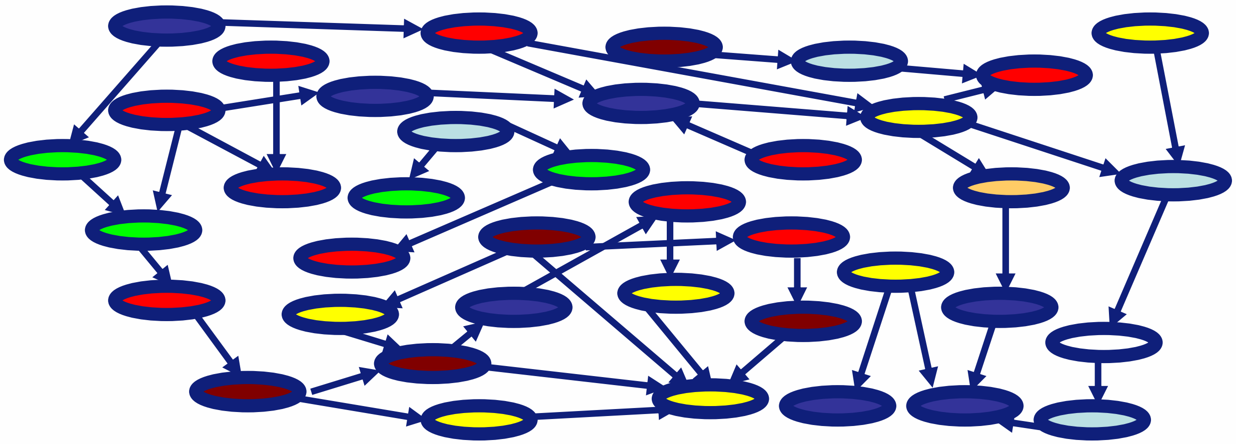
Tel +442380597682

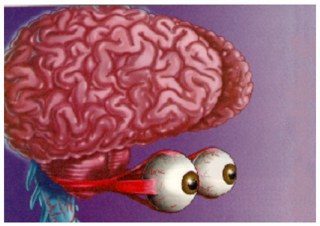
Fax +442380592865

Research
Fluid Dynamics
Aerodynamics
Design and Testing Technology
Biological Sciences Domain
Image and Vision Computing
Networks and Distributed Systems

CoP
Wendy Hall
DH Sleeman
DR Robertson
Stephen Harris
Hugh Glaser
M Eisenstadt
mkw
E Motta
Kieron O'Hara
W Hall
A Tate
Ian Millard
Les Carr
Y Wilks

SW as Enabler of Intelligent Behaviour

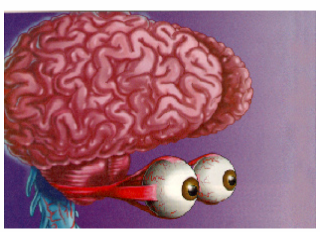




Intelligent Behaviour

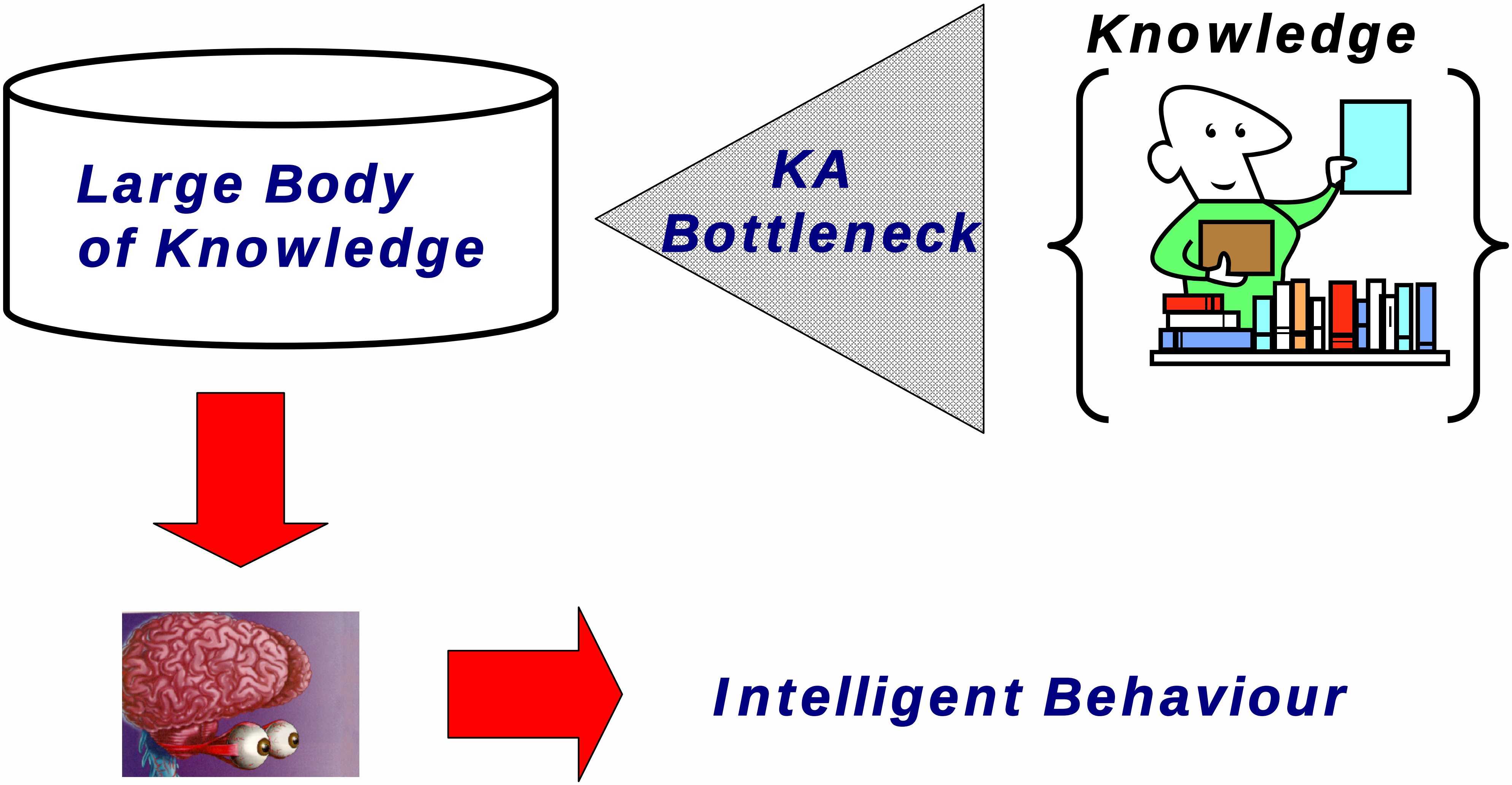
Intelligence as a function of possessing domain knowledge

Large Body of Knowledge



Intelligent Behaviour

The Knowledge Acquisition Bottleneck

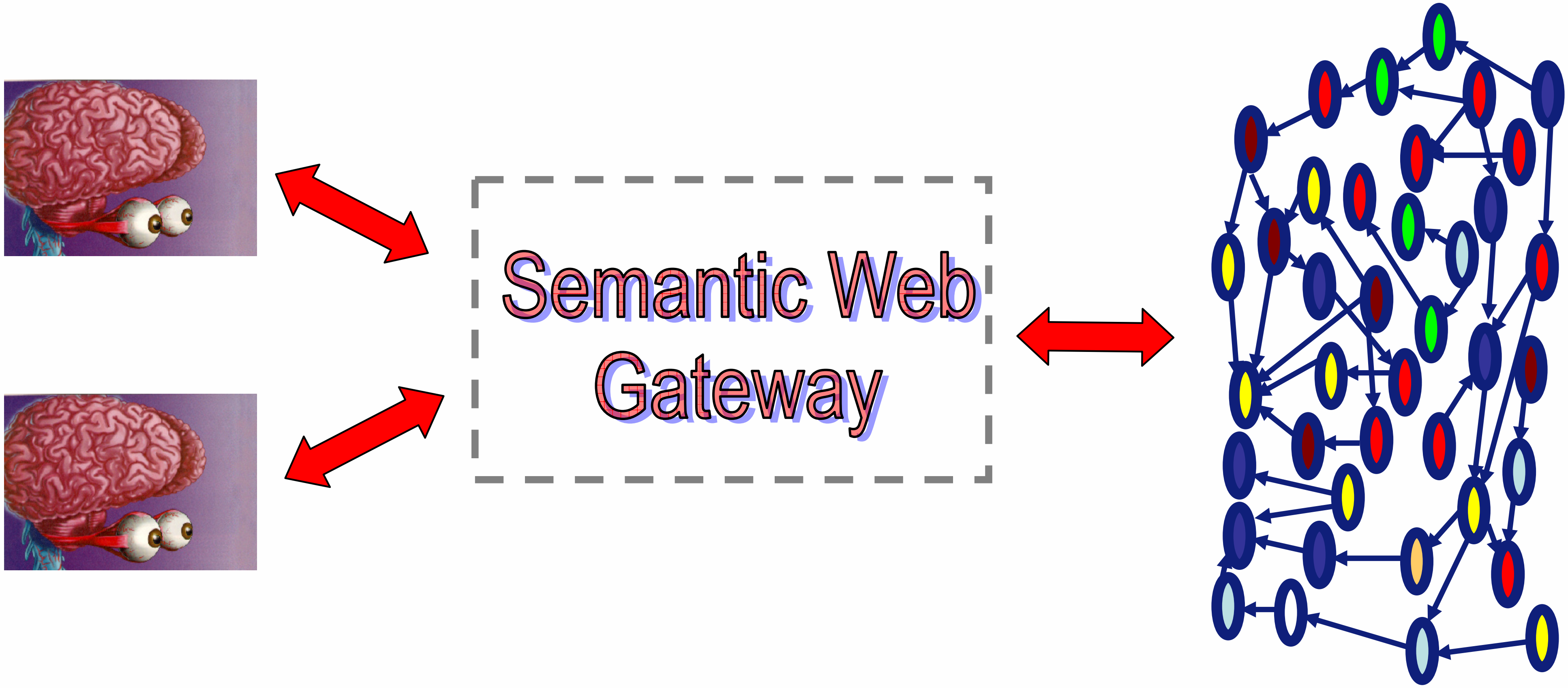


Thesis #1

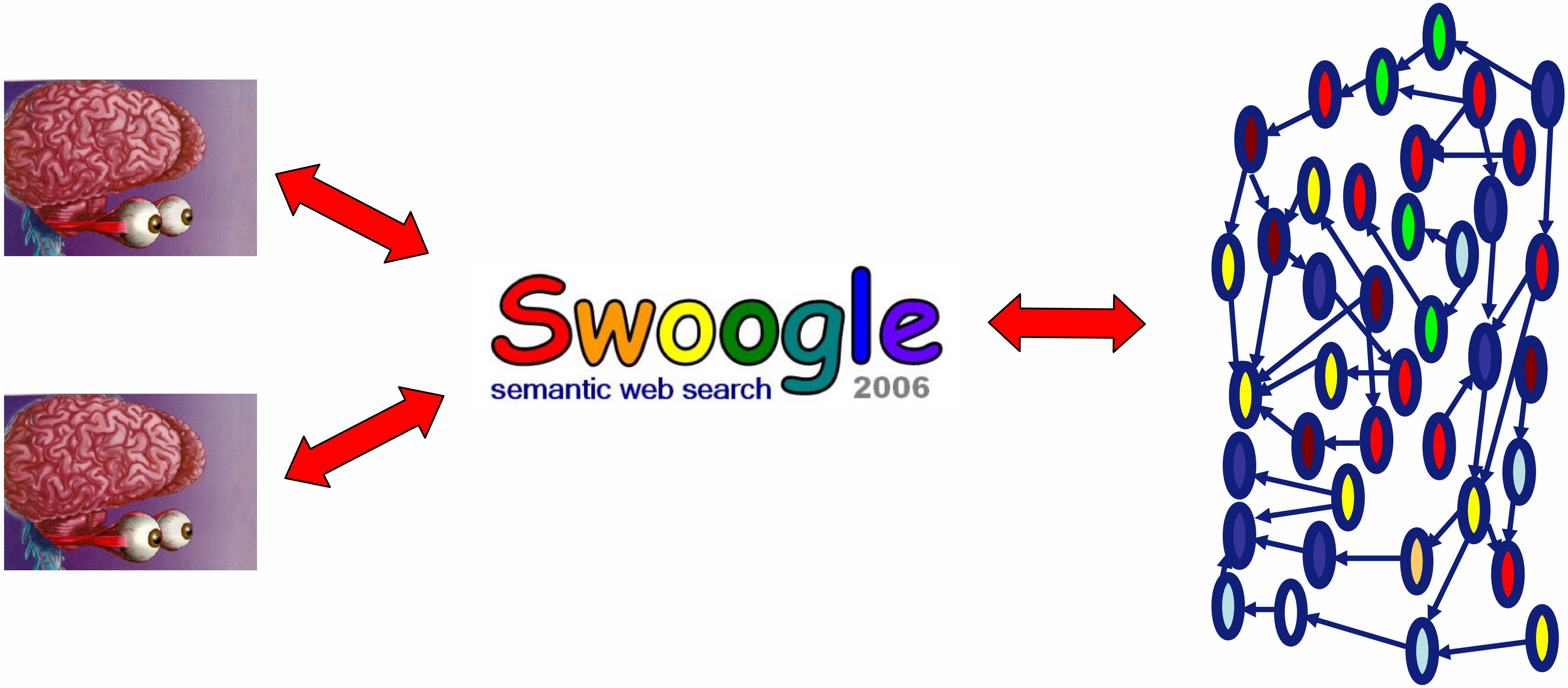
The SW may well provide a solution to one of the classic AI challenges: how to acquire and manage large volumes of knowledge to develop truly intelligent problem solvers and address the brittleness of traditional KBS

Infrastructure for 2G SW Applications

Architecture of NGSW Apps



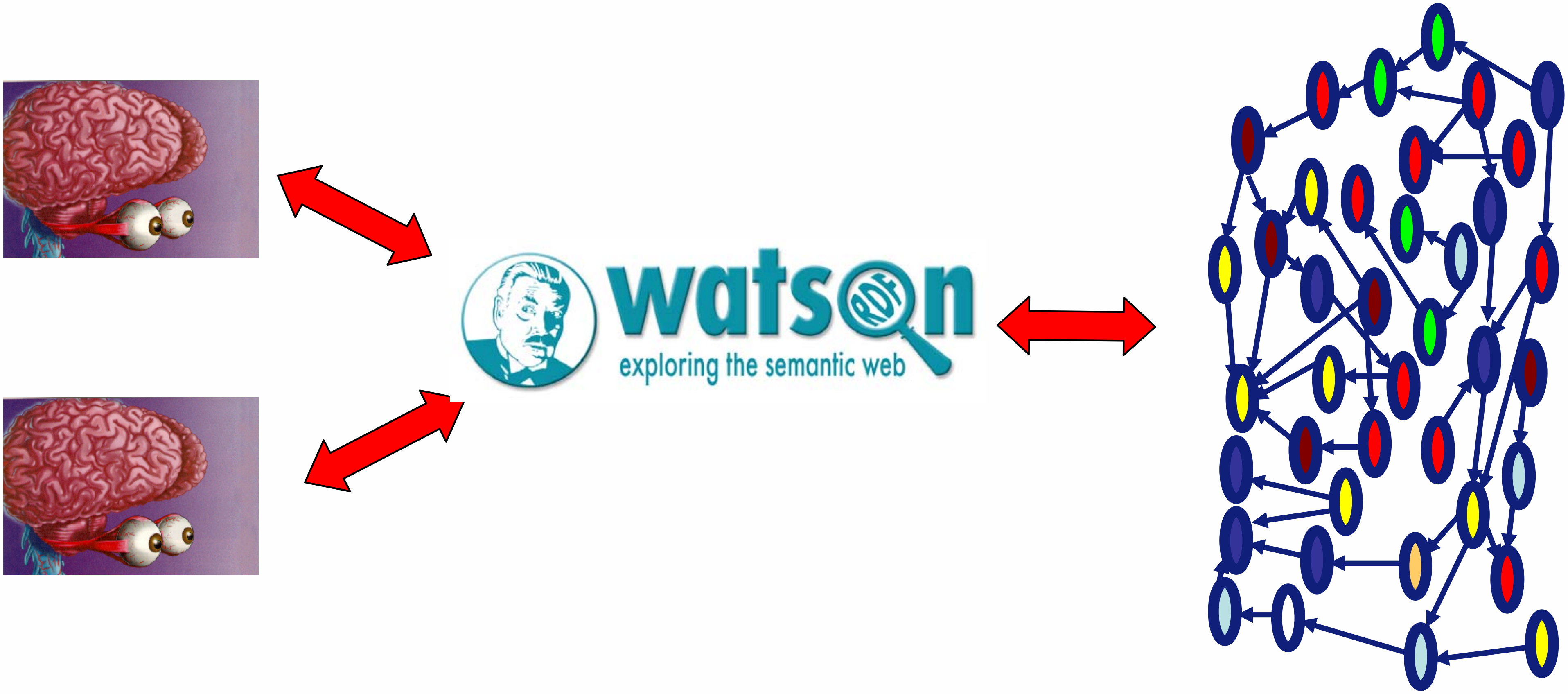
Current Gateway to the Semantic Web



Limitations of Swoogle

- Limited quality control mechanisms
  - Many ontologies are duplicated
- Limited Query/Search mechanisms
  - Only keyword search; no distinction between types of elements
  - No support for formal query languages (such as SPARQL)
- Limited range of ontology ranking mechanisms
  - Swoogle only uses a 'popularity-based' one
- Limited API
- No support for ontology modularization

A New Gateway to the Semantic Web

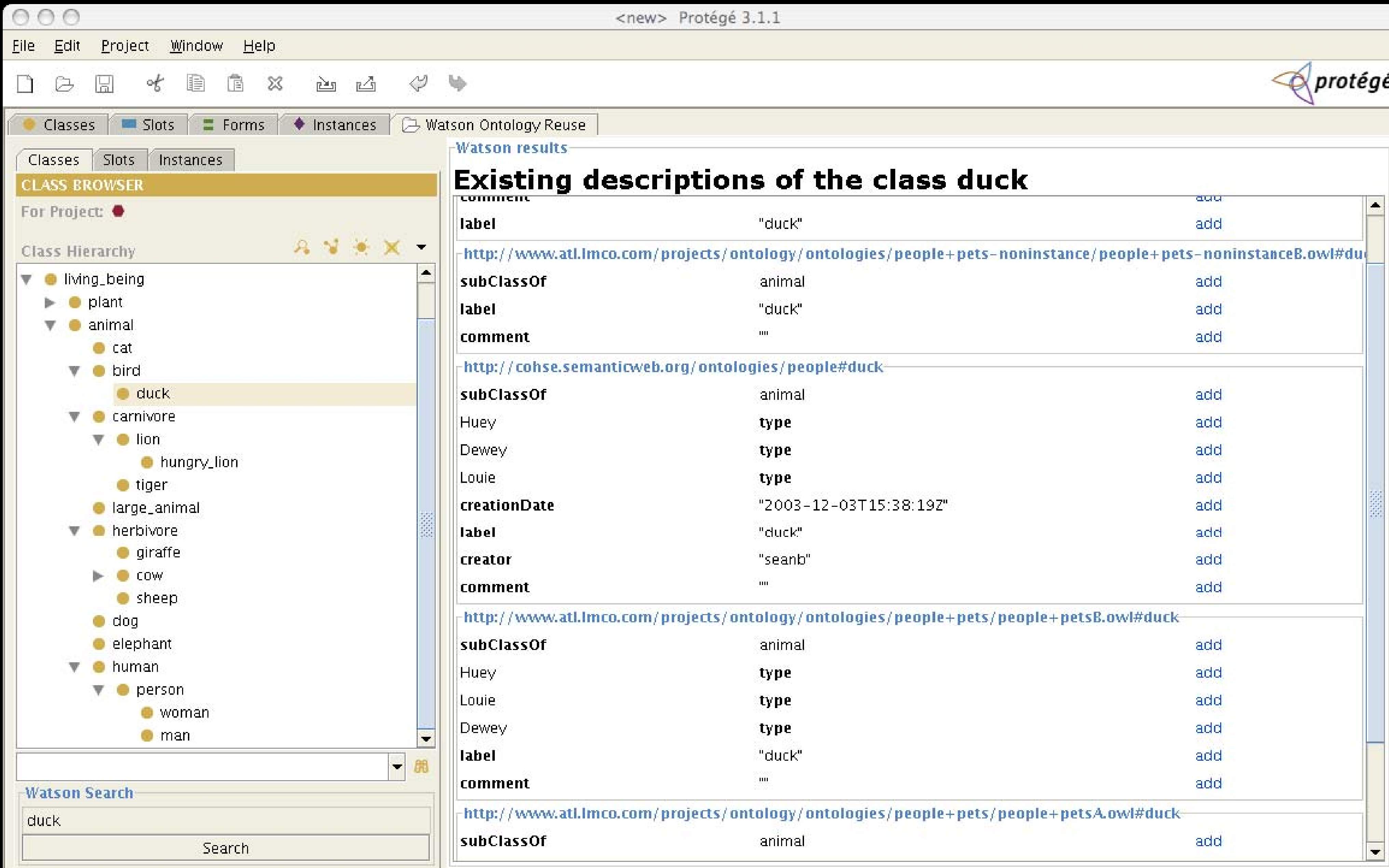


http://watson.kmi.open.ac.uk

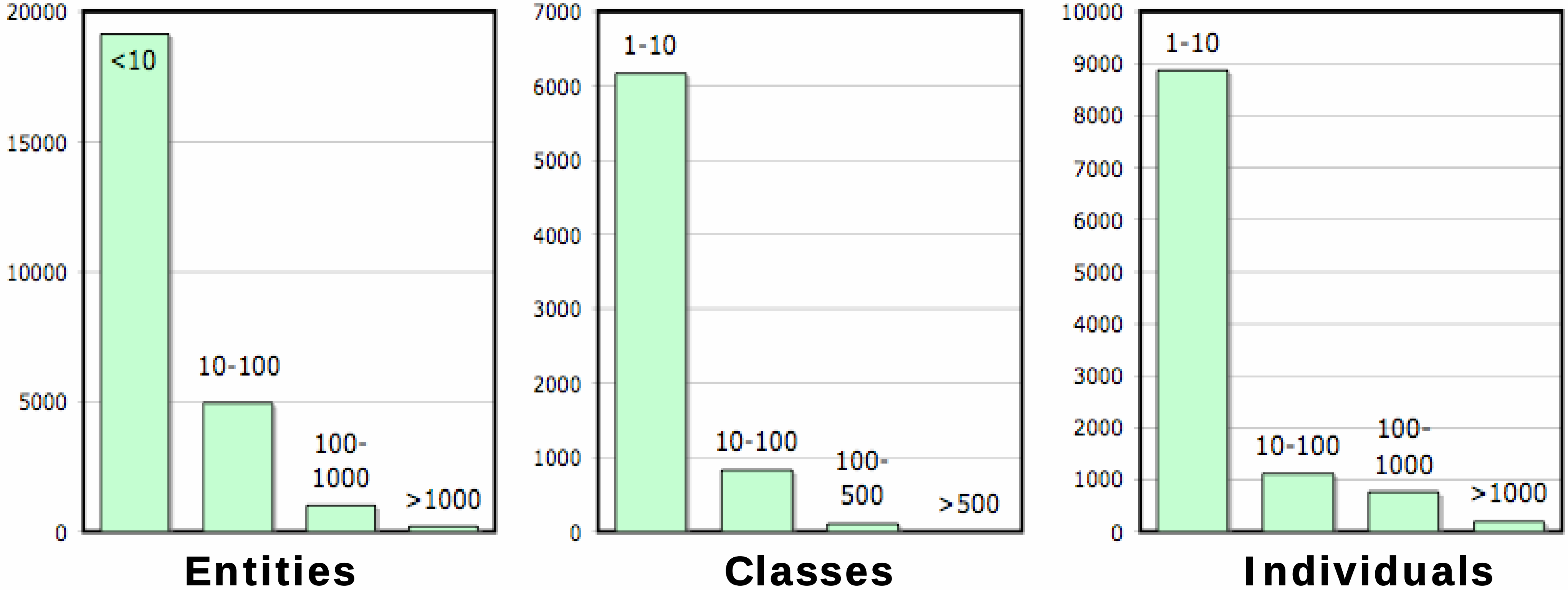
watson

exploring the semantic web

- Sophisticated quality control mechanism
  - Detects duplications
  - Fixes obvious syntax problems
    - E.g., duplicated ontology IDs, namespaces, etc..
- Structures ontologies in a network
  - Using relations such as: *extends, inconsistentWith, duplicates*
- Provides sophisticated API
- Supports formal queries (SPARQL)
- Supports a variety of ontology ranking mechanisms
- Modularization support
- Plug-ins for Protégé and NeOn Toolkit (both under devpt.)
- Very cool logo!


<new> Protégé 3.1.1
File
Edit
Project
Window
Help
protégé
Classes
Slots
Forms
Instances
Watson Ontology Reuse
Classes
Slots
Instances
CLASS BROWSER
For Project:
Class Hierarchy
living_being
plant
animal
cat
bird
duck
carnivore
lion
hungry_lion
tiger
large_animal
herbivore
giraffe
cow
sheep
dog
elephant
human
person
woman
man
Watson Search
duck
Search
Watson results
Existing descriptions of the class duck
label "duck" add
http://www.atl.lmco.com/projects/ontology/ontologies/people+pets-noninstance/people+pets-noninstanceB.owl#du
subClassOf animal add
label "duck" add
comment "" add
http://cohse.semanticweb.org/ontologies/people#duck
subClassOf animal add
Huey type add
Dewey type add
Louie type add
creationDate "2003-12-03T15:38:19Z" add
label "duck" add
creator "seanb" add
comment "" add
http://www.atl.lmco.com/projects/ontology/ontologies/people+pets/people+petsB.owl#duck
subClassOf animal add
Huey type add
Louie type add
Dewey type add
label "duck" add
comment "" add
http://www.atl.lmco.com/projects/ontology/ontologies/people+pets/people+petsA.owl#duck
subClassOf animal add

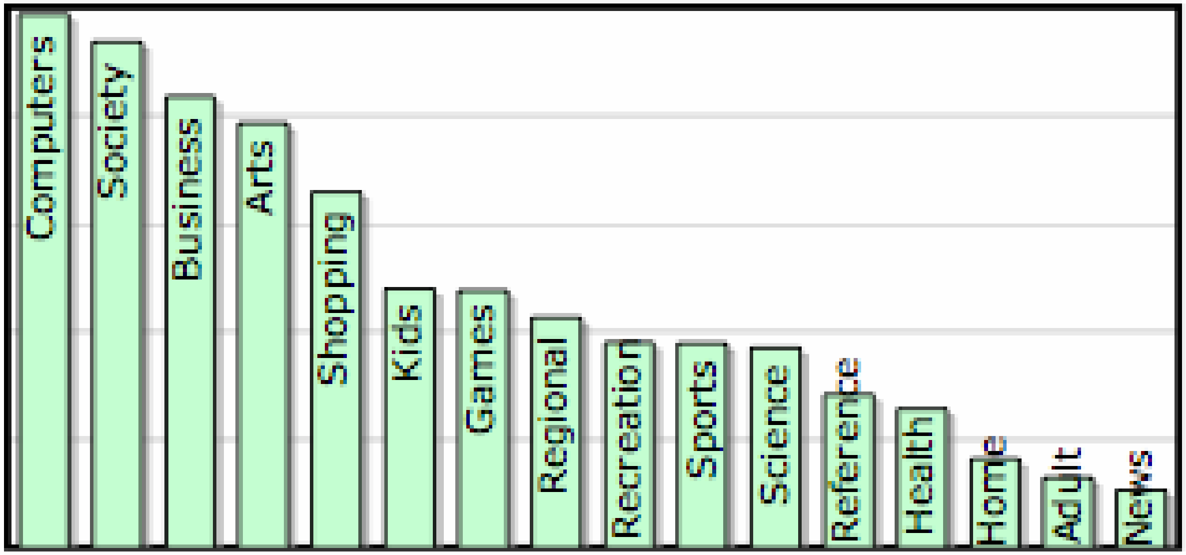
Charting the SW



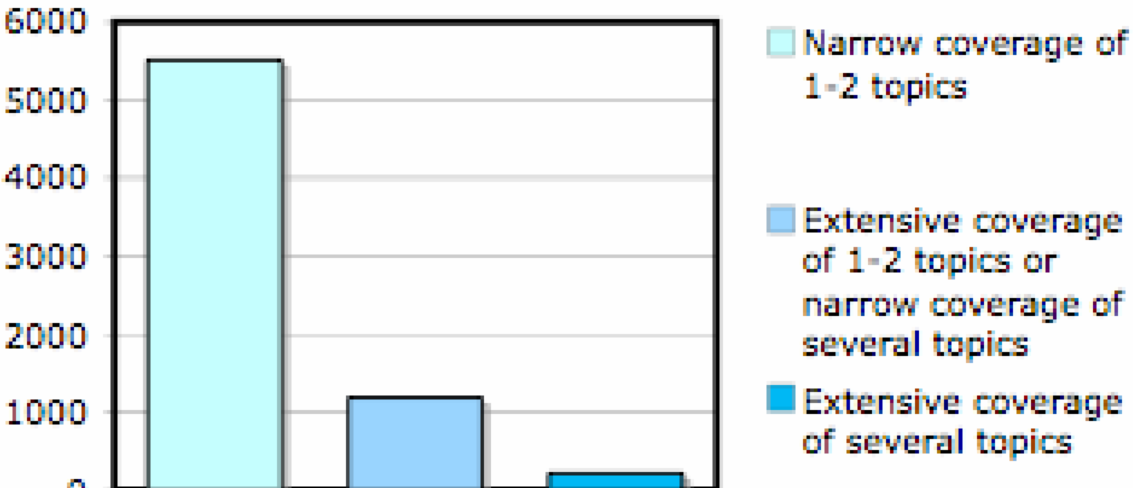
Distribution of SW documents according to the number of entities, classes and individuals

- –SW is characterized by a large number of small documents and a small number of large ones
- –This is true for both ontological knowledge (classes) and factual data (individuals)

Domain Coverage on the SW



Distribution of documents in the 16 top categories of DMOZ



Distribution of the documents according to their coverage

- Great variety: Some topics are almost not covered (e.g. Adult), while some are over represented (e.g. Society, Computers)
- As we can expect, a large number of narrow coverage documents and a small number of large coverage ones.

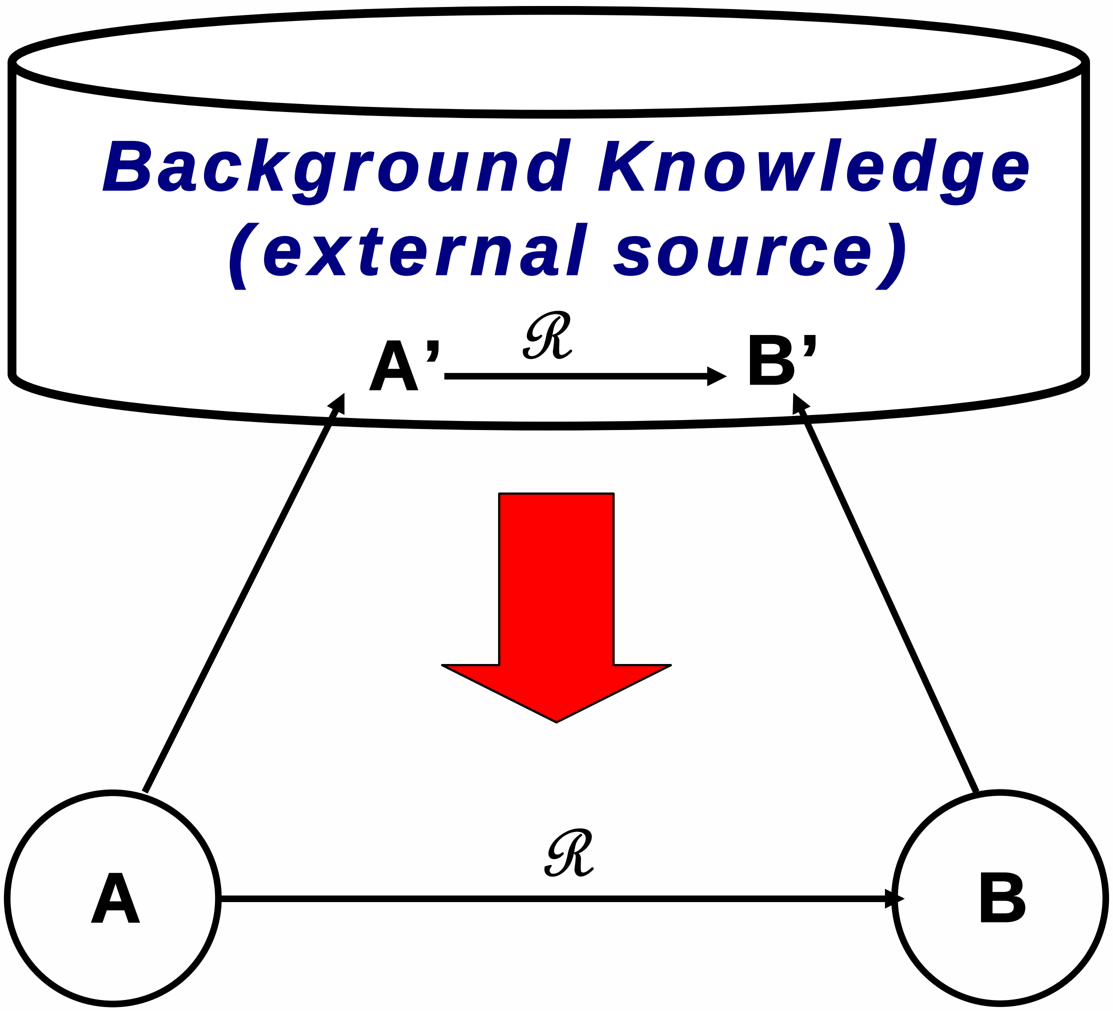
Density of the online knowledge

Measures	Value
Number of classes	161 264
Number of properties	76 350
Number of individuals	984 526
Number of sub-class relations	106 729
P-density (average number of properties per class)	**0.47**
H-density (average number of super-classes per class)	**0.66**
I-density (average number of instances per class)	**6.1**

- Usage of URIs for ontologies: lack of clear recommendation!
  - Most of the ontologies do not declare their URI
  - URI duplication and reuse:
    - Different versions of an ontology having the same URI (e.g. http: //lsdis.cs.uga.edu/proj/semdis/testbed/ used 4 times for 4 different versions, all available)
    - Mistaken use of a well known namespace (e.g. http://www.w3.org/2002/07/owl used as the URI of ontologies)
    - Default URI given by the ontology editor (e.g. http://a.com/ontology, the default URI in the OWL plugin of Protégé, used more than 20 times for ontologies having nothing to do together).

Example #2: Ontology Matching

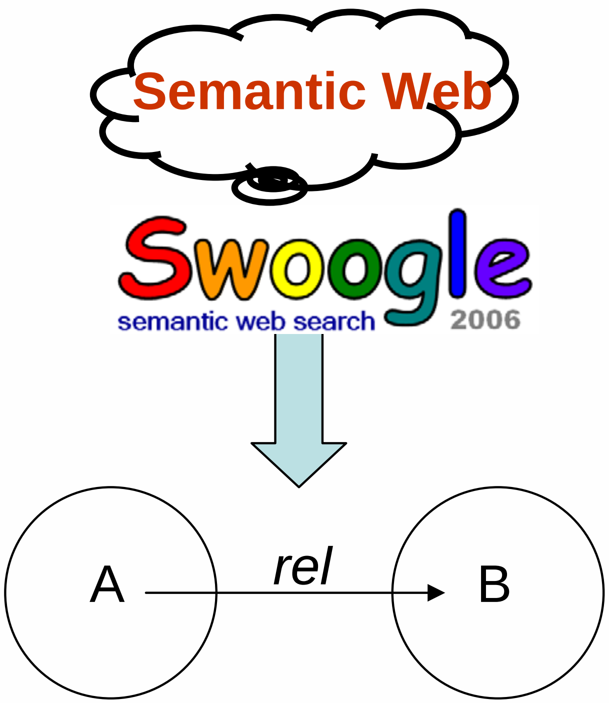
New paradigm: use of background knowledge



External Source = SW

Proposal:

- rely on online ontologies (Semantic Web) to derive mappings
- ontologies are ***dynamically*** discovered and combined



Does not rely on any pre-selected knowledge sources.

M. Sabou, M. d'Aquin, E. Motta, "*Using the Semantic Web as Background Knowledge in Ontology Mapping*", Ontology Mapping Workshop, ISWC'06. *Best Paper Award*

Examples

Ex1: $Chicken$ **Vs.** $Food$

$Chicken \subseteq Poultry$ **(midlevel-onto)**

$Poultry \subseteq Food$ **(Tap)**

(r1) ⇒ $Chicken \subseteq Food$

(Same results for Duck, Goose, Turkey)

Ex2: Ham **Vs.** $Food$

$Ham \subseteq Meat$ **(pizza-to-go)**

$Meat \subseteq Food$ **(SUMO)**

(r1) ⇒ $Ham \subseteq Food$

Ex3: Ham **Vs.** $Seafood$

$Ham \subseteq Meat$ **(pizza-to-go)**

$Meat \perp Seafood$ **(wine.owl)**

(r3) ⇒ $Ham \perp Seafood$

Large Scale Evaluation

Matching AGROVOC (16k terms) and NALT(41k terms)

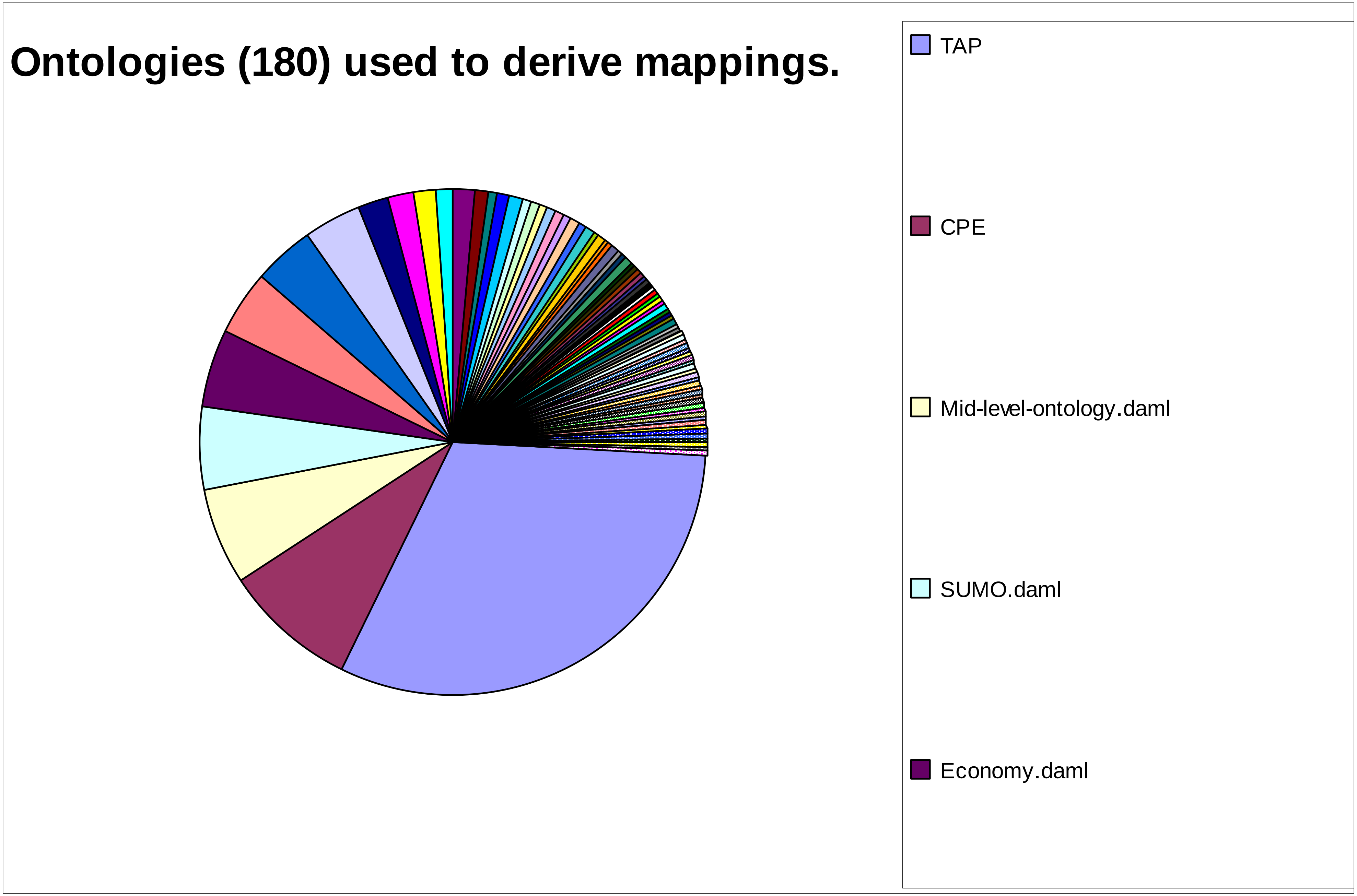
	Nr.	Examples
Subclass ($\xrightarrow{\sqsubseteq}$)	1477	$Lamb \xrightarrow{\sqsubseteq} Sheep$, $Soap \xrightarrow{\sqsubseteq} Detergent$, $Asbestos \xrightarrow{\sqsubseteq} Pollutant$ $Oasis \xrightarrow{\sqsubseteq} Ecosystem$, $RAM \xrightarrow{\sqsubseteq} ComputerEquipment$
SuperClass ($\xrightarrow{\sqsupseteq}$)	1857	$Shop \xrightarrow{\sqsupseteq} Supermarket$, $Spice \xrightarrow{\sqsupseteq} BlackPepper$, $Valley \xrightarrow{\sqsupseteq} Canyon$ $Infrastructure \xrightarrow{\sqsupseteq} Highway$, $Storm \xrightarrow{\sqsupseteq} Tornado$, $Rock \xrightarrow{\sqsupseteq} Crystal$
Disjoint ($\xrightarrow{\perp}$)	229	$Fluid \xrightarrow{\perp} Solid$, $Fluid \xrightarrow{\perp} Gas$, $Pond \xrightarrow{\perp} River$, $Plant \xrightarrow{\perp} Animal$ $Newspaper \xrightarrow{\perp} Journal$, $Fruit \xrightarrow{\perp} Vegetable$, $Female \xrightarrow{\perp} Male$
Total	3563	

(derived from 180 different ontologies)

Evaluation: 1600 mappings, two teams

Overall performance comparable to best in class (over 70%)

M. Sabou, M. d'Aquin, W.R. van Hage, E. Motta, "*Improving Ontology Matching by Dynamically Exploring Online Knowledge*". In Press



Ontologies (180) used to derive mappings.
TAP
CPE
Mid-level-ontology.daml
SUMO.daml
Economy.daml

Thesis #2

Using the SW to provide dynamically background knowledge to tackle the Agrovoc/NALT mapping problem provides the first ever test case in which the SW, viewed as a large scale heterogeneous resource, has been successfully used to address a real-world problem

<table>
<tr><th rowspan="2">Error Type</th><th rowspan="2">Nr./ %</th><th colspan="5">Examples</th></tr>
<tr><th>AGROVOC Concept</th><th>Labels</th><th>Rel.</th><th>NALT Concept</th><th>Labels</th></tr>
<tr><td rowspan="3">Anchor</td><td rowspan="3">114, 53%</td><td>c_6443</td><td>Rams, Tups</td><td>$\sqsubseteq$</td><td>memory</td><td>memory</td></tr>
<tr><td colspan="5">O_1:ram $\sqsubseteq$ O_1:memory</td></tr>
<tr><td colspan="5">O_1 = <code>http://www.arches.uga.edu/~gonen/qos_bilal.owl</code></td></tr>
<tr><td rowspan="3">Subsumption as generic relation</td><td rowspan="3">40, 18%</td><td>c_3954</td><td>Irrigation</td><td>$\sqsubseteq$</td><td>agriculture</td><td>agriculture</td></tr>
<tr><td colspan="5">O_1:Irrigation $\sqsubseteq$ O_1:SoilCultivation $\sqsubseteq$ O_1:Agriculture</td></tr>
<tr><td colspan="5">O_1 = <code>http://sweet.jpl.nasa.gov/ontology/human_activities.owl</code></td></tr>
<tr><td rowspan="3">Subsumption as part-whole</td><td rowspan="3">16, 7%</td><td>c_666</td><td>Asia</td><td>$\sqsupseteq$</td><td>Iran</td><td>Iran</td></tr>
<tr><td colspan="5">O_1:Asia $\sqsupseteq$ O_1:WestAsia $\sqsupseteq$ O_1:Iran</td></tr>
<tr><td colspan="5">O_1 = <code>http://islab.hanyang.ac.kr/damls/Country.daml</code></td></tr>
<tr><td rowspan="3">Subsumption as role</td><td rowspan="3">11, 5%</td><td>c_11091</td><td>Garlic</td><td>$\sqsubseteq$</td><td>ingredients</td><td>ingredients</td></tr>
<tr><td colspan="5">O_1:garlic $\sqsubseteq$ O_1:vegetable $\sqsubseteq$ O_1:ingredient</td></tr>
<tr><td colspan="5">O_1 = <code>http://cvs.sourceforge.net/viewcvs.py/instancestore/instancestore/ontologies/Attic/pizza9.daml?rev=1.2</code></td></tr>
<tr><td rowspan="3">Inaccurate labeling</td><td rowspan="3">12, 5%</td><td>c_1693</td><td>Coal</td><td>$\sqsubseteq$</td><td>industry</td><td>industry</td></tr>
<tr><td colspan="5">O_1:coal $\sqsubseteq$ O_1:industry</td></tr>
<tr><td colspan="5">O_1 = <code>http://www.aifb.uni-karlsruhe.de/WBS/meh/mapping/data/russia1a.rdf</code></td></tr>
<tr><td rowspan="3">Different View</td><td rowspan="3">12, 5%</td><td>c_2943</td><td>Fishes</td><td>$\sqsupseteq$</td><td>lobsters</td><td>lobsters</td></tr>
<tr><td colspan="5">O_1:Fish $\sqsupseteq$ O_1:MarineInvertebrate $\sqsupseteq$ O_1:Crustacean $\sqsupseteq$ O_1:Lobster</td></tr>
<tr><td colspan="5">O_1 = <code>http://139.91.183.30:9090/RDF/VRP/Examples/tap.rdf</code></td></tr>
</table>

Thesis #3

The claim that the information on the SW is of poor quality and therefore not useful to support intelligent problem solving is a myth not supported by concrete experience:

Our experience in the NALT/Agrovoc ontology matching benchmark problem shows that without any particularly intelligent filter, the info available on the SW already allows a 85% theoretical precision for our algorithm, well beyond the performance of any other ontology matching algorithm

Conclusions

- SW provides an unprecedented opportunity to build a new generation of intelligent systems, able to exploit large scale background knowledge
- The large scale background knowledge provided by the SW may address one of the fundamental premises (and holy grails) of AI
- The SW is not an aspiration: it is a concrete technology that is already in place today and is steadily becoming larger and more robust
- This new scenario opens up new opportunities, however we also need new methods and tools to support the life-cycle of the envisaged applications, which is the goal of the NeOn project
- The applications shown in this talk provide an initial taster of the kind of opportunities the SW will provide for intelligent problem solving

KNOWLEDGE MEDIA

KMi

INSTITUTE