A Research Programme for the Semantic Web

Enrico Motta, Sofia Angeletou, Claudio Baldassarre, Mathieu D’Aquin, Martin Dzbor, Laurian Gridinoc, Davide Guidi, Vanessa Lopez, Marta Sabou

Knowledge Media Institute
The Open University
Milton Keynes, UK
This talk presents a number of projects, which are part of an integrated effort at exploring the possibilities opened by the Semantic Web, viewed as a domain-independent, large scale supplier of formally encoded background knowledge, with respect to enabling intelligent problem solving. We call the resulting applications: **Next Generation Semantic Web Applications**
Organization of the Talk

• The Semantic Web
• The Semantic Web in the context of AI research
• Next Generation Semantic Web Applications
  – What are they?
  – Why are they different from 1st generation SW Applications?
• Examples
  – Ontology Matching
  – Integrating Web2.0 and SW
  – Semantic Web Browsing
  – Question Answering
• Conclusions
The collection of all formal, machine processable, web accessible, ontology-based statements (semantic metadata) about web resources and other entities in the world, expressed in a knowledge representation language based on an XML syntax (e.g., OWL, DAML, DAML+OIL, RDF, etc...)
People

Members [76]

Affiliates [5]

Alumni [43]

[People/All Members]

[People/All Members]

[A][B][C][D][E][F][G][H][I][J][K][L][M][N][O][P][Q][R][S][T][U][V][W][X][Y][Z]

Paul Alexander
IT & Systems Manager [info] [homepage] [email] [RDF/XML]

Carlo Allocca
Intern [info] [email] [RDF/XML]

Sofia Angeletou
Research Student [info] [email] [RDF/XML]

Michelle Suzanne Bachler
Project Officer [info] [homepage] [email] [RDF/XML]

David Bainbridge
Visiting Researcher [info] [homepage] [RDF/XML]

Claudio Baldassarre
Research Assistant [info] [homepage] [email] [RDF/XML]
Charting the web
Swoogle's Statistics of the Semantic Web

### Swoogle Today

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>admin_dt</td>
<td>2007-03-25 23:58:45</td>
<td>Datetime Watched</td>
</tr>
<tr>
<td>url_total</td>
<td>5,249,633</td>
<td>Number of URLs being discovered</td>
</tr>
<tr>
<td>url_pinged</td>
<td>3,219,485</td>
<td>Number of URLs being pinged</td>
</tr>
<tr>
<td>total_swd</td>
<td>2,048,446</td>
<td>Number of Semantic Web Documents (regardless of embedded or containing some errors) be confirmed.</td>
</tr>
<tr>
<td>total_swd_strict</td>
<td>1,183,514</td>
<td>Number of error-free pure Semantic Web Documents</td>
</tr>
<tr>
<td>total_swd_embed</td>
<td>688,233</td>
<td>Number of documents (except SWDs, PDF, and JPEG) embedding Semantic Web Data</td>
</tr>
<tr>
<td>triple_total</td>
<td>368,620,315</td>
<td>Number of triples could be parsed from all Semantic Web Documents.</td>
</tr>
</tbody>
</table>
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]

Shenley Church End
Shenley
United Kingdom
populated place
N 52° 1' 0" W 0° 47' 0"
52.01667 / -0.78333

<rdf:Description rdf:about=""/>
<mediapro:People>
  <rdf:Bag>
    <rdf:li>Jim Hendler</rdf:li>
    <rdf:li>Enrico</rdf:li>
  </rdf:Bag>
</mediapro:People>
The Rise of Semantics

Semantic Technologies Center

Semantic Technologies are designed to extend the capabilities of ink enterprises to be networked in meaningful ways. The adoption of Consortium (W3C) standards like XML, RDF (Resource Description Framework), Ontology Language) serve as foundation technologies to advancing technologies.

Oracle Spatial 10g introduces the industry's first open, scalable, geospatial data model. RDF triples are persisted, indexed and quickly searched. The 10g RDF database ensures that application developers benefit from secure semantic applications. Application areas include:

Finding and Exploiting Value in Semantic Technologies on the Web

9 May 2007

David W. Cearley, Whit Andrews, Nicholas Gall

DataPatrol is a new monthly database to design and de

Revolutionary online protection

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

The future of the Web is Semantic

Level: Introductory
The SW today has already reached a level of scale good enough to make it a very useful source of knowledge to support intelligent applications.

In other words: the Semantic Web is no longer an aspiration but a reality.

The availability of such large scale amounts of formalised knowledge is unprecedented in the history of AI.
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.
Any mechanically embodied intelligent process will be comprised of structural ingredients that we as external observers naturally take to represent a propositional account of the knowledge that the overall process exhibits, and independent of such external semantic attribution, play a formal but causal and essential role in engendering the behaviour that manifests that knowledge.

Brian Smith, 1982
Intelligence as a function of possessing domain knowledge

Large Body of Knowledge

Intelligent Behaviour
The Knowledge Acquisition Bottleneck

Large Body of Knowledge

KA Bottleneck

Knowledge

Intelligent Behaviour
Building Large Knowledge-Based Systems

Representation and Inference in the Cyc Project

Douglas B. Lenat
R. V. Guha
SW as Enabler of Intelligent Behaviour

Intelligent Behaviour
Our research programme is to contribute to the development of this large-scale web of data and develop a new generation of web applications able to exploit it to provide intelligent functionalities.
Overview: NR Shadbolt

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
Issue: Semantic Web Infrastructure
Current Gateway to the Semantic Web

Swoogle

semantic web search 2006
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
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)

Variety of ontology ranking mechanisms

Modularization support

Plug-ins for Protégé and NeOn Toolkit (under devpt.)

Very cool logo!
Existing descriptions of the class duck

- label: "duck"
- http://www.atlaimco.com/projects/ontology/ontologies/people+pets+noninstance/people+pets+noninstance.owl#duck
- subclassOf: animal
- http://cohse.semanticweb.org/ontologies/people#duck
- creationDate: "2003-12-03T15:38:19Z"
- label: "duck"
- creator: "seanb"
- http://www.atlaimco.com/projects/ontology/ontologies/people+pets/people+pets.owl#duck
- subclassOf: animal
Examples of Next Generation Semantic Web Applications
Example #1: Ontology Matching
Label similarity methods
- e.g., Full_Professor = FullProfessor

Structure similarity methods
- Using taxonomic/property related information
New paradigm: use of background knowledge

Background Knowledge (external source)
Aleksovski et al. EKAW’06

- Map (anchor) terms into concepts from a richly axiomatized domain ontology
- Derive a mapping based on the relation of the anchor terms

Assumes that a suitable (rich, large) domain ontology (DO) is available.
van Hage et al. ISWC’05
• rely on Google and an online dictionary in the food domain to extract semantic relations between candidate terms using IR techniques

Does not rely on a rich Domain Ont,

Precision increases significantly if domain specific sources are used:
50% - Web;
75% - domain texts.
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*
How to combine online ontologies to derive mappings?
Strategy 1 - Definition

Find ontologies that contain equivalent classes for A and B and use their relationship in the ontologies to derive the mapping.

For each ontology use these rules:

\[ A' \equiv B' \Rightarrow A \equiv B \]
\[ A' \subseteq B' \Rightarrow A \subseteq B \]
\[ A' \supseteq B' \Rightarrow A \supseteq B \]
\[ A' \perp B' \Rightarrow A \perp B \]

These rules can be extended to take into account indirect relations between A’ and B’, e.g., between parents of A’ and B’:

\[ A' \subseteq C \land C \perp B' \Rightarrow A' \perp B' \]
Strategy 1 - Examples

Semantic Web

**SR-16**

- **FAO_Agrovoc**

**Tap**

- **Beef**
  - **Food**
  - **MeatOrPoultry**
  - **RedMeat**
  - **Beef**

**ka2.rdf**

**Researcher**

- **AcademicStaff**

**ISWC**

- **SWRC**

**Swoogle** 2006
**Principle:** If no ontologies are found that contain the two terms then combine information from multiple ontologies to find a mapping.

**Details:**
(1) Select all ontologies containing A’ equiv. with A  
(2) For each ontology containing A’:

(a) if $A' \subseteq C$ find relation between C and B.  
(b) if $A' \supseteq C$ find relation between C and B.

(r1) $A' \subseteq C \land C \subseteq B \Rightarrow A \subseteq B$  
(r2) $A' \subseteq C \land C \equiv B \Rightarrow A \equiv B$  
(r3) $A' \subseteq C \land C \perp B \Rightarrow A \perp B$  
(r4) $A' \supseteq C \land C \supseteq B \Rightarrow A \supseteq B$  
(r5) $A' \supseteq C \land C \equiv B \Rightarrow A \supseteq B$
Strategy 2 - Examples

Ex1: *Chicken* vs. *Food*

\[
\begin{align*}
\text{Chicken} & \subseteq \text{Poultry} \quad \text{(midlevel-onto)} \\
\text{Poultry} & \subseteq \text{Food} \quad \text{(Tap)}
\end{align*}
\]

\[
\text{Ex1: } \text{Chicken} \subseteq \text{Food}
\]

(Same results for Duck, Goose, Turkey)

Ex2: *Ham* vs. *Food*

\[
\begin{align*}
\text{Ham} & \subseteq \text{Meat} \quad \text{(pizza-to-go)} \\
\text{Meat} & \subseteq \text{Food} \quad \text{(SUMO)}
\end{align*}
\]

\[
\text{Ex2: } \text{Ham} \subseteq \text{Food}
\]

Ex3: *Ham* vs. *Seafood*

\[
\begin{align*}
\text{Ham} & \subseteq \text{Meat} \quad \text{(pizza-to-go)} \\
\text{Meat} & \perp \text{Seafood} \quad \text{(wine.owl)}
\end{align*}
\]

\[
\text{Ex3: } \text{Ham} \perp \text{Seafood}
\]
Matching AGROVOC (16k terms) and NALT(41k terms)

<table>
<thead>
<tr>
<th></th>
<th>Nr.</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subclass (⊆)</td>
<td>1477</td>
<td>Lamb ⊆ Sheep, Soap ⊆ Detergent, Asbestos ⊆ Pollutant</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oasis ⊆ Ecosystem, RAM ⊆ ComputerEquipment</td>
</tr>
<tr>
<td>SuperClass (⊇)</td>
<td>1857</td>
<td>Shop ⊇ Supermarket, Spice ⊇ BlackPepper, Valley ⊇ Canyon</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Infrastructure ⊇ Highway, Storm ⊇ Tornado, Rock ⊇ Crystal</td>
</tr>
<tr>
<td>Disjoint (⊥)</td>
<td>229</td>
<td>Fluid ⊥ Solid, Fluid ⊥ Gas, Pond ⊥ River, Plant ⊥ Animal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Newspaper ⊥ Journal, Fruit ⊥ Vegetable, Female ⊥ Male</td>
</tr>
<tr>
<td>Total</td>
<td>3563</td>
<td></td>
</tr>
</tbody>
</table>

(derived from 180 different ontologies)

Evaluation: 1600 mappings, two teams
Overall performance comparable to best in class

Ontologies (180) used to derive mappings.
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.
Example #2: 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

All time most popular tags

08 africa amsterdam animals architecture art august australia 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 de dog england europe faq family festival fl 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
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
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 dog england europe fall family festival fl florida flower flowers food france friends fun garden geotagged germany girl graffiti green halloween hawaii hong 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 200
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} \rightarrow <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
Concept and relation identification

Folksonomy

Clustering

Analyze co-occurrence of tags

Co-occurrence matrix

Cluster tags

Cluster 1, Cluster 2, ..., Cluster n

2 “related” tags

Find mappings & relation for pair of tags

SW search engine

Wikipedia

Google

<concept, relation, concept>

Pre-processing

Tags

Clean tags

Group similar tags

Filter infrequent tags

Concise tags

Remaining tags?

Yes

No

END
• **Scope:** Subsets of *Flickr* and *del.icio.us* tags.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># entries</td>
<td># tags</td>
</tr>
<tr>
<td>del.icio.us</td>
<td>19,605</td>
<td>89,978</td>
</tr>
<tr>
<td>Flickr</td>
<td>49,087</td>
<td>167,130</td>
</tr>
</tbody>
</table>

• **Pre-processing (thresholds):**
  - To be “similar”, Levenshtein $\geq 0.83$;
  - A tag has to occur at least **10** times.
**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}**

• Approach proven to be feasible and promising.
• However...
  – Assumptions in initial experiments (e.g., single ontology coverage for pairs of tags; focus on classes, clustering-based approach, etc..) too restrictive
  – Swoogle is too limited to support a fully automated approach
    ➔ we are now using Watson for the current experiments
  – Integration with SW-enabled ontology matching algorithm is essential to improve term matching
Example #3: Semantics-Enhanced Web Browsing
ANSES - [Multimedia and Information Systems]

Automatic News Summarization and Extraction System

"Watch the news while I was away and tell me what happened." This project combines a Video scene change algorithm, with the current text segmentation and summarization techniques to build an automatic news summarization and extraction system. Television broadcast news are captured both in Video/Audio format with the accompanying subtitles in text format. News stories are identified, extracted from the Video, and summarized in a short paragraph which reduces the amount of information into a manageable size. Individual news video clips can be retrieved effectively by a combination of Video and text, using a reversed indexed search engine to provide distilled information such as a summarized version of the original text and highlights important key words in the text.

Participants: Marcus Pickering, Lawrence Wong

Technology Champion: Stefan Ruger [email] [info]

URL: http://kmi.open.ac.uk/technologies/anses/
ANSES - [Multimedia and Information Systems]

Automatic News Summarization and Extraction System

'Watch the news while I was away and tell me what happened.'

This project combines a video scene change algorithm, with the current text segmentation and summarization techniques to build an automatic news summarization and extraction system. Television broadcasts of news are captured both in video/audio format with the accompanying subtitles in text format. News stories are identified, extracted from the video, and summarized in a short paragraph which reduces the amount of information into a manageable size. Individual news clips can be retrieved effectively by a combination of video and text, using a reverse indexed search engine to provide distilled information such as a summarized version of the original text and highlights important keywords in the text.

Participants: Marcus Pickering, Lawrence Wong

Technology Champion: Stefan Rüger [email] [info]

URL: http://kmi.open.ac.uk/technologies/anses/
ANSES - [Multimedia and Information Systems]

Automatic News Summarization and Extraction System

'Watch the news while I was away and tell me what happened.'

This project combines a visual scene change algorithm, with the current text segmentation and summarization techniques to build an automatic news summarization and extraction system. Telephone broadcast news are captured both in Video Avivo format with identifiers and in a XML-RDF format. The XML-RDF version of the original text and highlights important key words in the text.

Participants: Marcus Pickering, Lawrence Wong

Technology Champion: Stefan Rieger [email] [info]

URL: http://kmi.open.ac.uk/technologies/anses/
ANSES - [Multimedia and Information Systems]

Automatic News Summarization and Extraction System

Watch the news while I was away and tell me what happened.

This project combines a video scene change algorithm, with the current text segmentation and summarization techniques to build an automatic news summarization and extraction system.

Televised broadcast news are captured both in video and audio format with a high-quality capture system. The video and audio signals are digitized and analyzed. The most important part is the scene change detection. A scene change is detected by analyzing the video signal. The captured audio signal is used to select the right sound for each scene.

Semantic Web Definitions

http://www.memento.org/data.basis.rdf

Video Type

http://www.diku.dk/heidelberg/kurs/ss05/ki/index.rdf

Subclass of

http://www.diku.dk/heidelberg/kurs/ss03/ki/Uebungen/Online/ontology.rdf#Television

creationDate

14:31:12 01.02.2002

Comment

Enabling news programs to broadcast visual images of stationary or moving objects
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.

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 examples shown in this talk provide an initial taste of the new generation of applications which will be made possible by the emerging Semantic Web.
• Ontology Mapping

• Integration of Web2.0 and Semantic Web

• Watson

