Semantic Technologies an overview

Marko Grobelnik
Marko.Grobelnik@ijs.si
Jozef Stefan Institute
Ljubljana, Slovenia

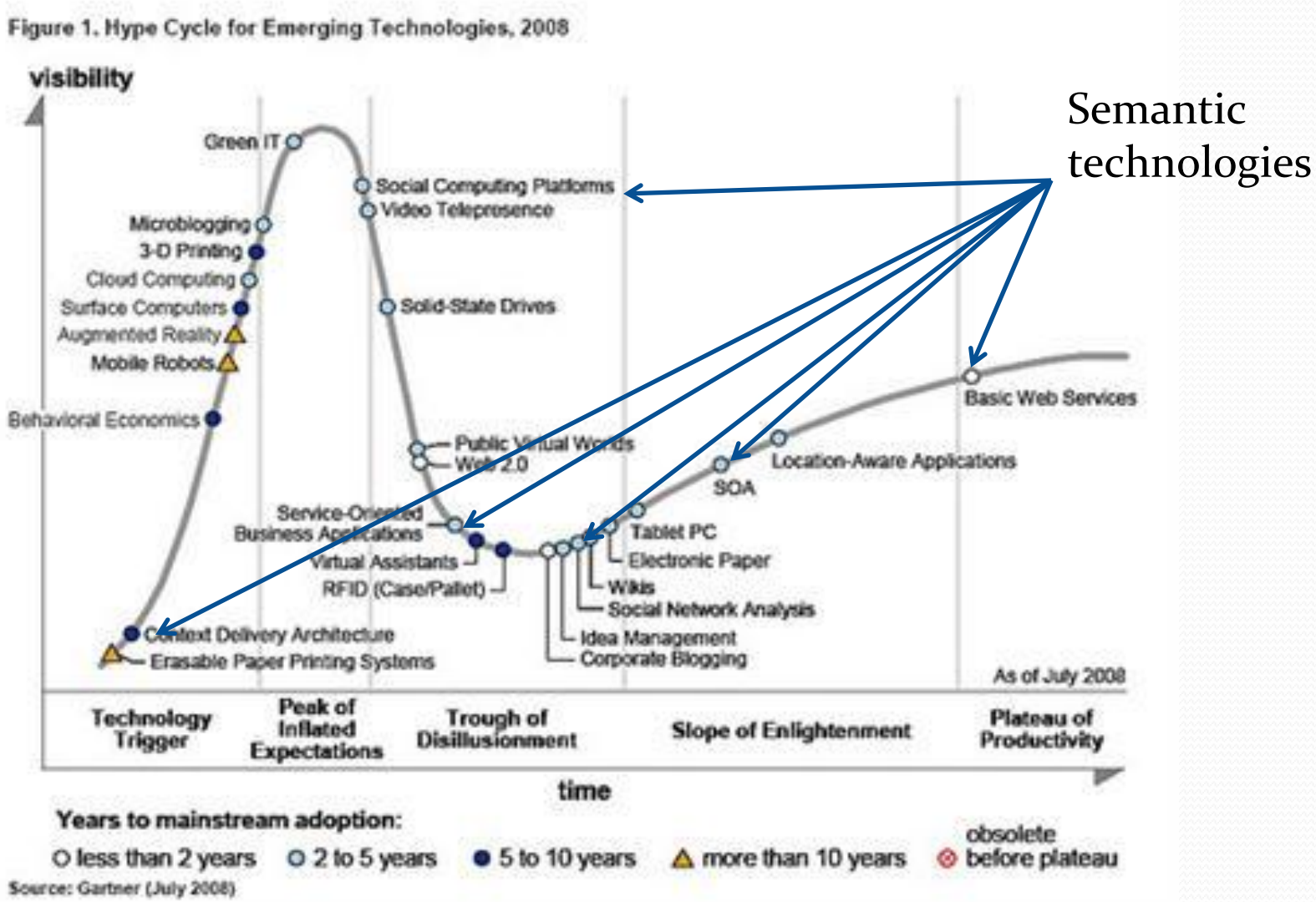
Outline

- Quick introduction
  - …what are semantic technologies?
- Gartner's hype curve
- Semantic Web Technology stack
- Web X.X
- Examples
  - Dealing with legacy relational databases
  - Dealing with legacy software
  - Contextualized search
  - Identifying news reporting bias
  - Common sense reasoning

What are semantic technologies?

- Semantic technologies are interdisciplinary set of technologies with the main goal to make information **interoperable**
- What are the three main "buzzwords"?
  - Semantic Web
  - Semantic Web Services
  - Web2.0
- …and related ones:
  - W3C, Social computing, Ontologies, … and many more

Where Semantic technologies fit into Gartner's hype-cycle



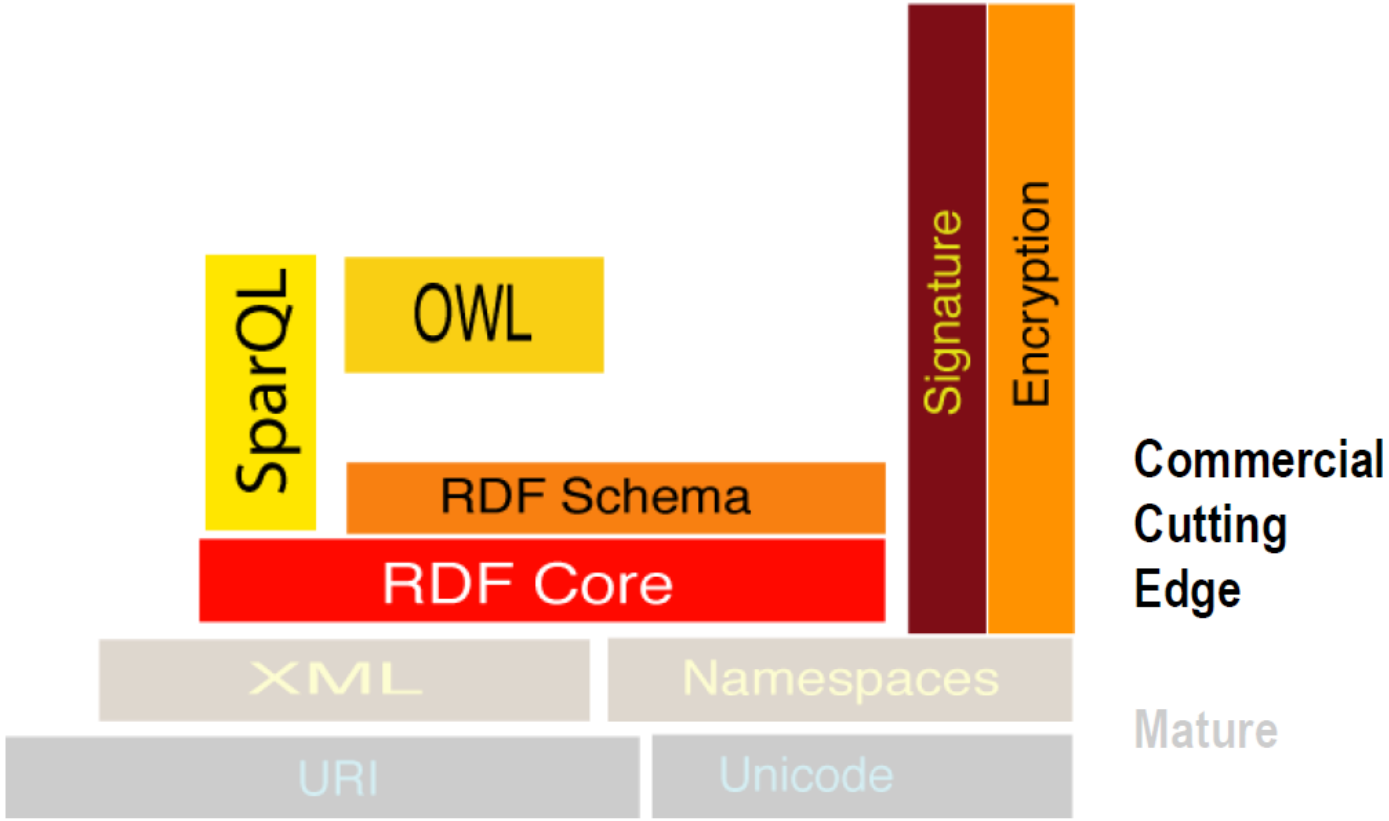
Semantic Web Technology stack

The Semantic Web in 2008

Mature

"The Famous Semantic Web Technology Stack"

The Semantic Web in 2008

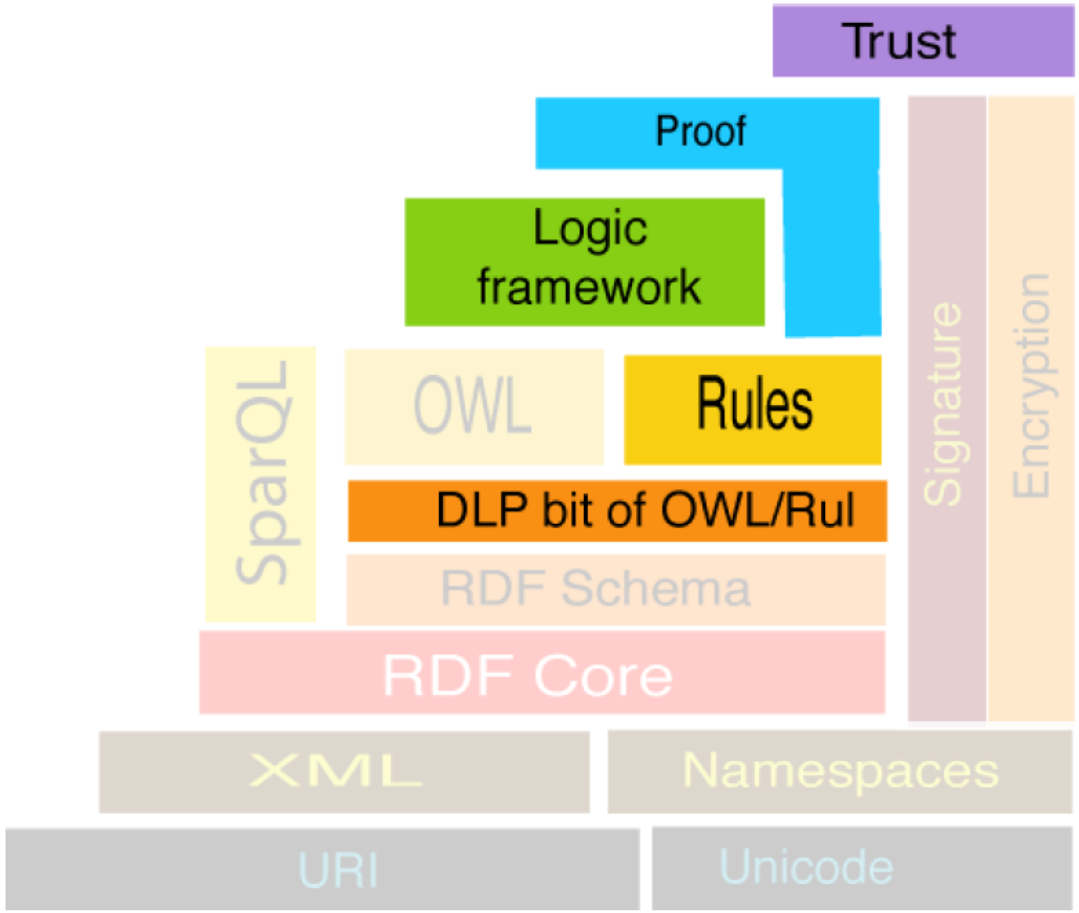


"The Famous Semantic Web Technology Stack"

Compiled by Mark Greaves

VULCAN

The Semantic Web in 2008



Active Research and Standards Activity

Commercial Cutting Edge

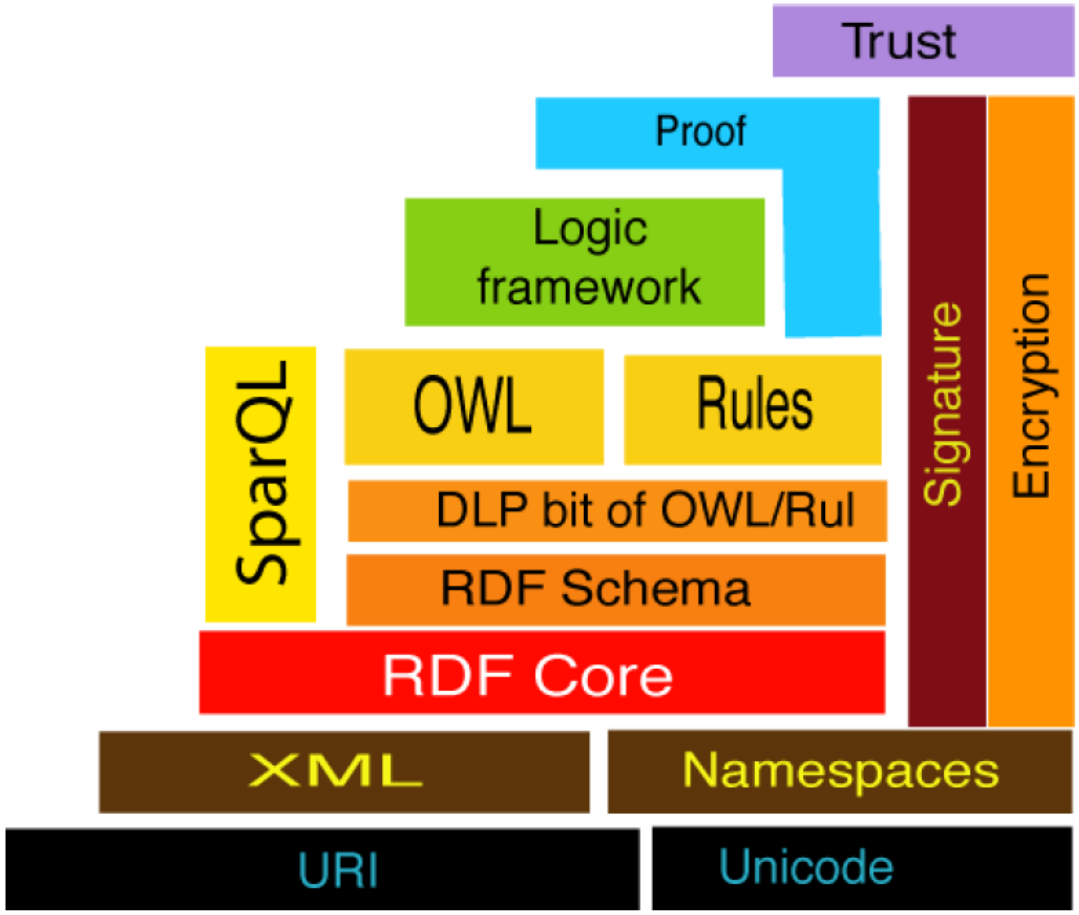
Mature

"The Famous Semantic Web Technology Stack"

Compiled by Mark Greaves

VULCAN

The Semantic Web in 2008



Active Research and Standards Activity

Commercial Cutting Edge

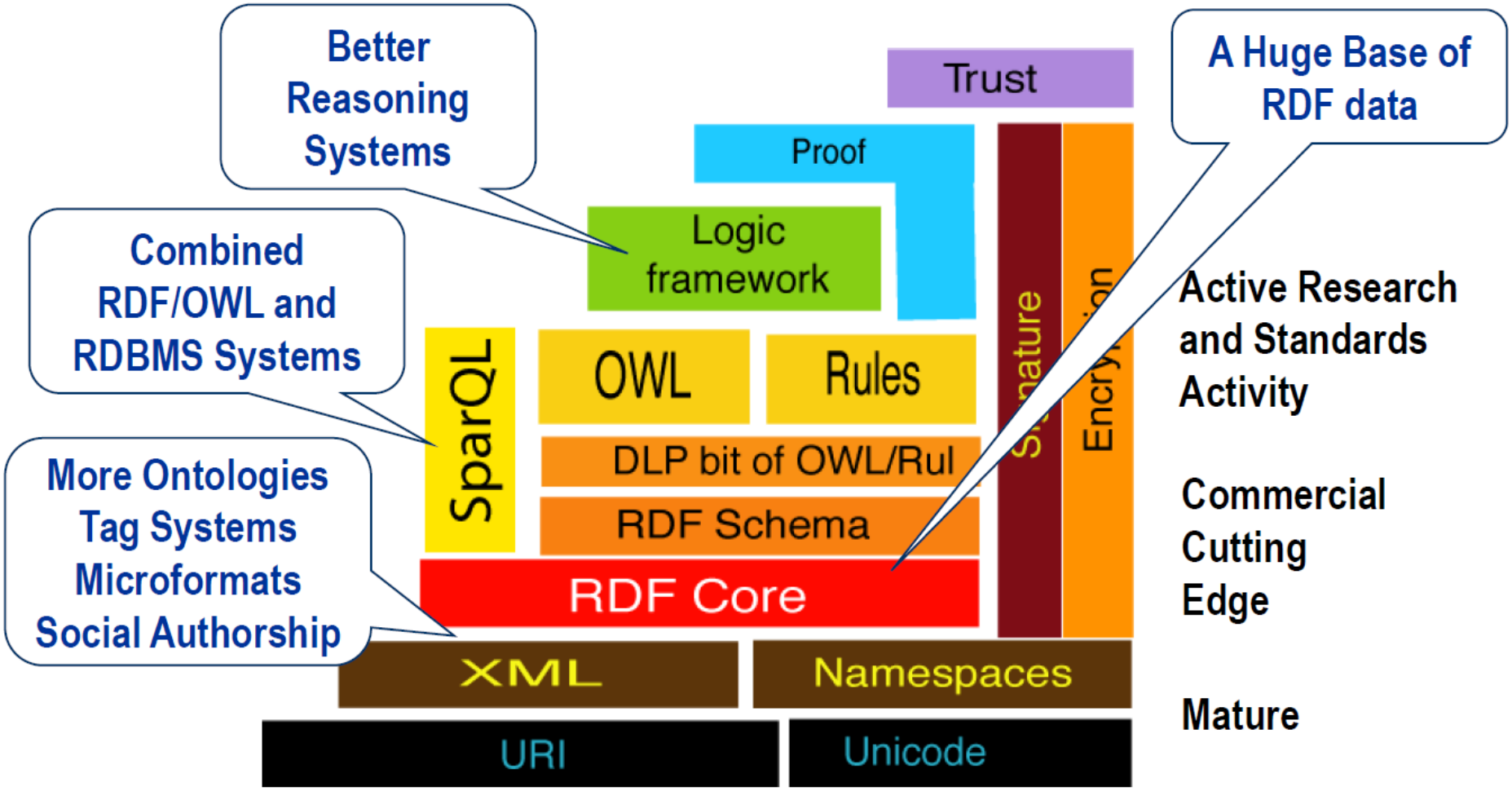
Mature

"The Famous Semantic Web Technology Stack"

Compiled by Mark Greaves

VULCAN

Completing the Semantic Web Picture



Other Technologies Impact the Semantic Web

The beautiful world of Web X.X

The beautiful world of Web X.X versions (…a trial to put all of them on one slide)

	Description	Technologies
Web 1.0	**Static** HTML pages (web as we first learned it)	HTML, HTTP
Web 1.5	**Dynamic** HTML content (web as we know it)	Client side (JavaScript, DHTML, Flash, …), server side (CGI, PHP, Perl, ASP/.NET, JSP, …)
Web 2.0	**Integration** on all levels, collaboration, sharing vocabularies (web as it is being sold)	weblogs, social bookmarking, social tagging, wikis, podcasts, RSS feeds, many-to-many publishing, web services, … URI, XML, RDF, OWL, …
Web 3.0	…adding **meaning** to semantics - AI dream revival (web as we would need it)	Closest area of a research would be "common sense reasoning" and the "Cyc system" (http://www.nytimes.com/2006/11/12/business/12web.html?ref=business)

Web 2.0 –is there any new quality?

- IMHO, with "Web 2.0" the Web community became **really aware** of the importance of the global collaborative work
  - …next step in globalization of the Web
  - **Bottom-up** "social networking" seems to nicely complement the traditional **top-down** schema design approaches

Visualization of Web 2.0 typical vocabulary (http://en.wikipedia.org/wiki/Image:Web20_en.png)

Web 2.0 – the current hype

Google search volume of “Web 2.0” vs. “semantic web” vs. “data mining”

…scale and dynamics of Web 2.0

- Per minute, there are:
  - 100 edits in Wikipedia (144K/day)
  - 200 tags in del.icio.us (288K/day)
  - 270 image uploads to flickr (388K/day)
  - 1100 blog entries (1.6M/day)

What about Web 4.0? ☺

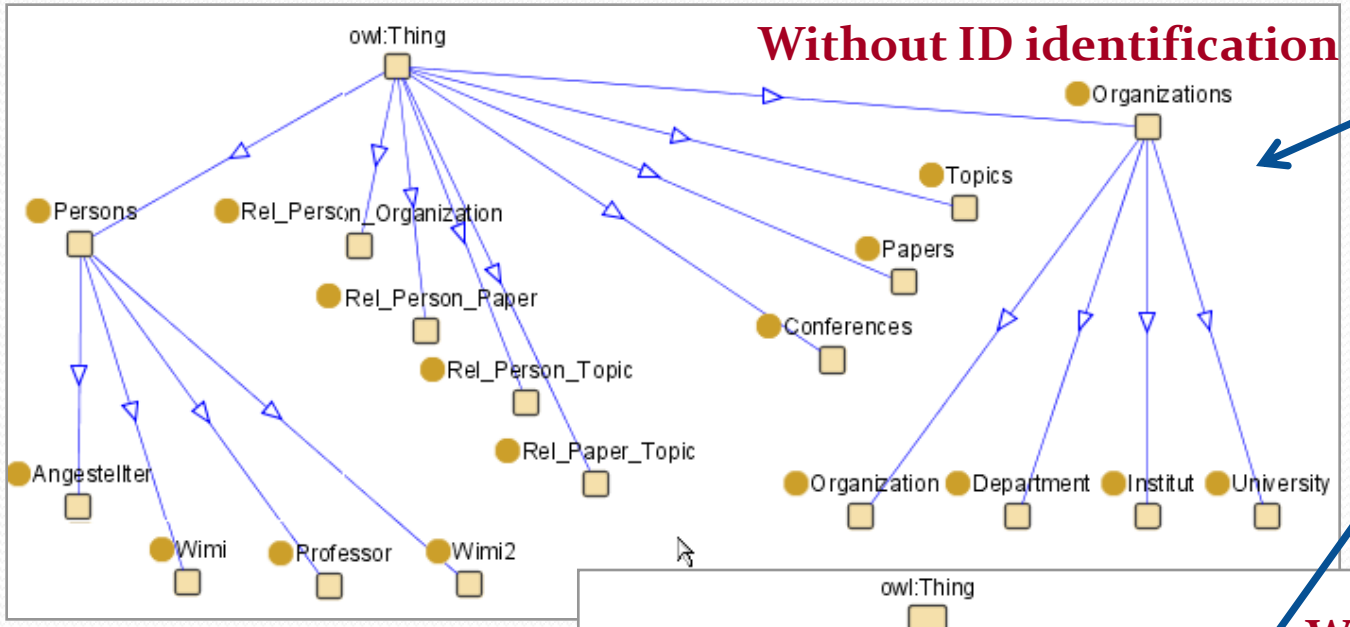
- Citation from some Intel blog:
  - "*…Web 4.0 is the impending state at which all information converges into a great ball of benevolent self-aware light, and solves every problem from world peace to …*" http://blogs.intel.com/it/2006/11/web_40_a_new_hype.html
- Ultimate stage in web development…
  - …will prevent Web 5.0 to happen since everything will be resolved already by Web 4.0.

Examples & Demos

Example: mining data models in legacy databases

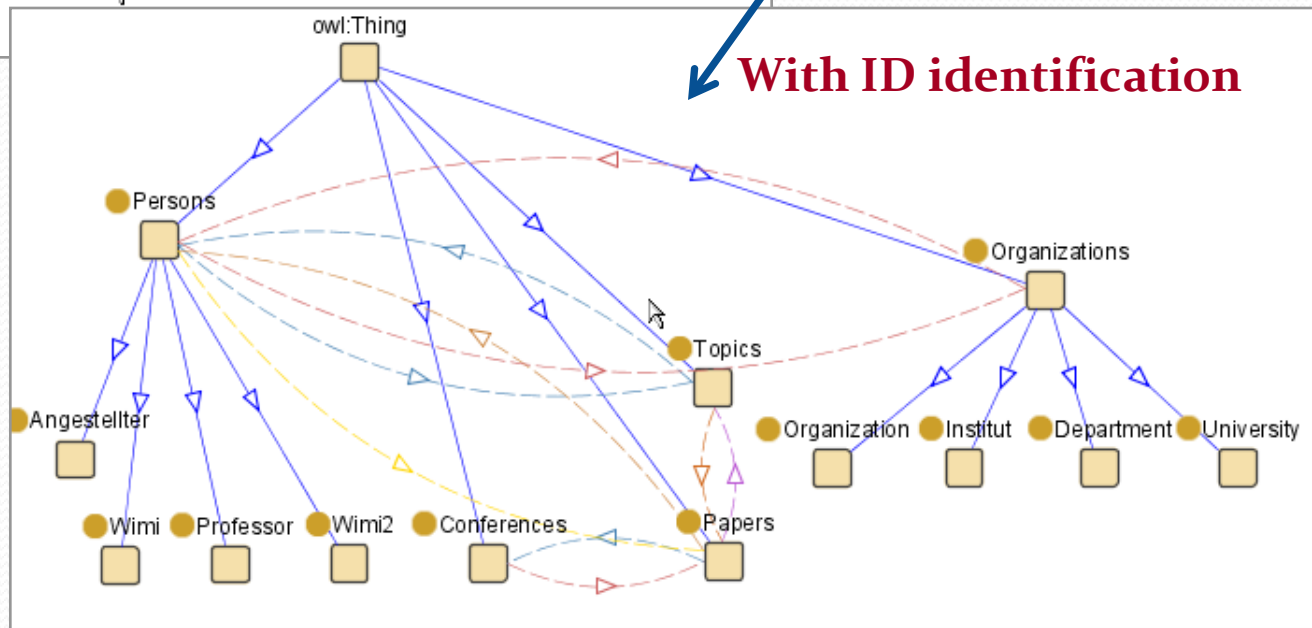
- Data models in relational databases are often not designed properly
  - …especially after many patches and many people being involved
- In the next example we show how a large relational database (~500 tables) from DESSAULT (airplane producer) was corrected with semiautomatic system

Example: finding hidden foreign-key relationships in large relational databases



Legacy database

Proposed relationships



Example: legacy software mining

- Software is as any other data source possible domain for analysis
- …in the following example we are mining large legacy software package GATE written in Java and present some alternative views

Software Data Sources

- Structured
  - Code samples
  - Web service usage logs
  - Source code
  - DB schemas …
- Unstructured
  - Web pages
  - User's/Reference manual
  - Tutorials, lectures, forums, newsgroups, etc.
  - Source code comments
  - DB content …

A Typical Java Class

Class comment

```
/** The format of Documents. Subclasses of DocumentFormat know about
  * particular MIME types and how to unpack the information in any
  * markup or formatting they contain into GATE annotations. Each MIME
  * type has its own subclass of DocumentFormat, e.g. XmlDocumentFormat,
  * RtfDocumentFormat, MpegDocumentFormat. These classes register themselves
  * with a static index residing here when they are constructed. Static
  * getDocumentFormat methods can then be used to get the appropriate
  * format class for a particular document.
  */
public abstract class DocumentFormat
extends AbstractLanguageResource implements LanguageResource{
```

```
/** The MIME type of this format. */
 private MimeType mimeType = null;
```

Field comment

Method comment

```
/**
  * Find a DocumentFormat implementation that deals with a particular
  * MIME type, given that type.
  * @param  aGateDocument this document will receive as a feature
  *                       the associated Mime Type. The name of the feature is
  *                       MimeType and its value is in the format type/subtype
  * @param  mimeType the mime type that is given as input
  */
 static public DocumentFormat getDocumentFormat(gate.Document aGateDocument,
                                                MimeType mimeType){
```

```
  } // getDocumentFormat(aGateDocument, MimeType)

} // class DocumentFormat
```

Creating a Document Network

DocumentFormat.class

```
/** The format of Documents. Subclasses of DocumentFormat know about
  * particular MIME types and how to unpack the information in any
  * markup or formatting they contain into GATE annotations. Each MIME
  * type has its own subclass of DocumentFormat, e.g. XmlDocumentFormat,
  * RtfDocumentFormat, MpegDocumentFormat. These classes register themselves
  * with a static index residing here when they are constructed. Static
  * getDocumentFormat methods can then be used to get the appropriate
  * format class for a particular document.
  */
public abstract class DocumentFormat
extends AbstractLanguageResource implements LanguageResource{
```

```
/** The MIME type of this format. */
 private MimeType mimeType = null;
```

```
/**
   * Find a DocumentFormat implementation that deals with a particular
   * MIME type, given that type.
   * @param  aGateDocument this document will receive as a feature
   *                       the associated Mime Type. The name of the feature is
   *                       MimeType and its value is in the format type/subtype
   * @param  mimeType the mime type that is given as input
   */
 static public DocumentFormat getDocumentFormat(gate.Document aGateDocument,
                                                MimeType mimeType){
```

```
 } // getDocumentFormat(aGateDocument, MimeType)

} // class DocumentFormat
```

DocumentFormat

Creating a Document Network

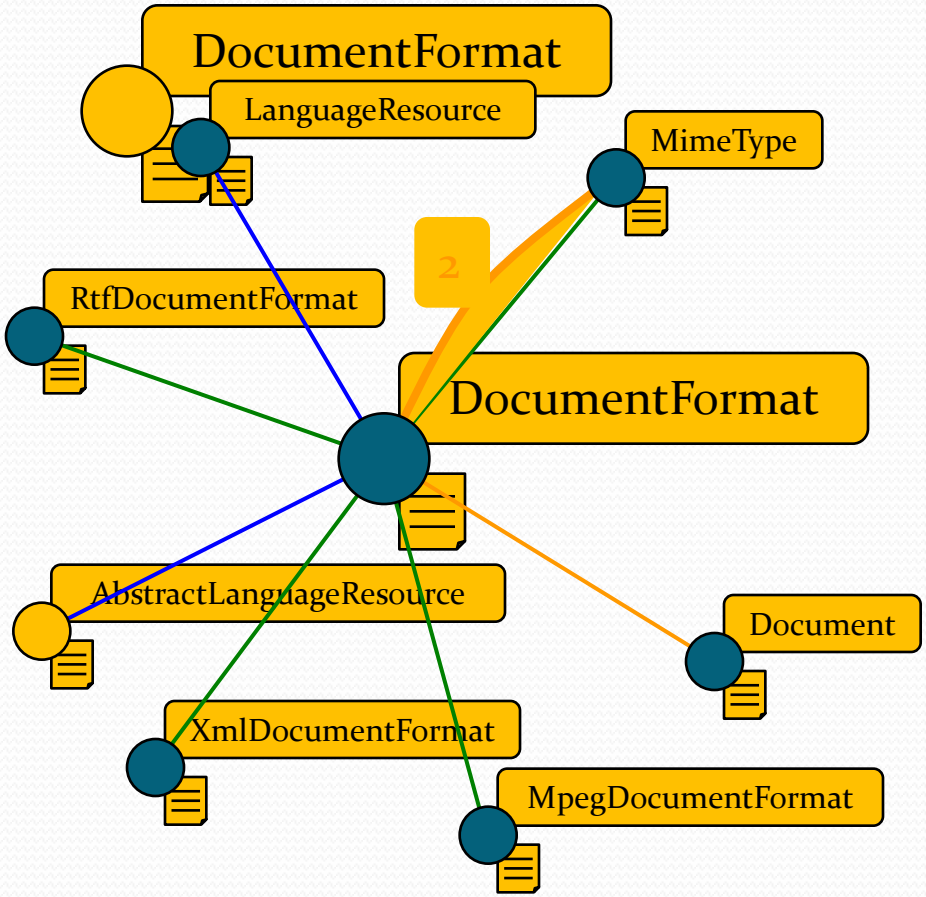
DocumentFormat.class

```
/** The format of Documents. Subclasses of DocumentFormat know about
  * particular MIME types and how to unpack the information in any
  * markup or formatting they contain into GATE annotations. Each MIME
  * type has its own subclass of DocumentFormat, e.g. XmlDocumentFormat,
  * RtfDocumentFormat, MpegDocumentFormat. These classes register themselves
  * with a static index residing here when they are constructed. Static
  * getDocumentFormat methods can then be used to get the appropriate
  * format class for a particular document.
  */
public abstract class DocumentFormat
extends AbstractLanguageResource implements LanguageResource{
```

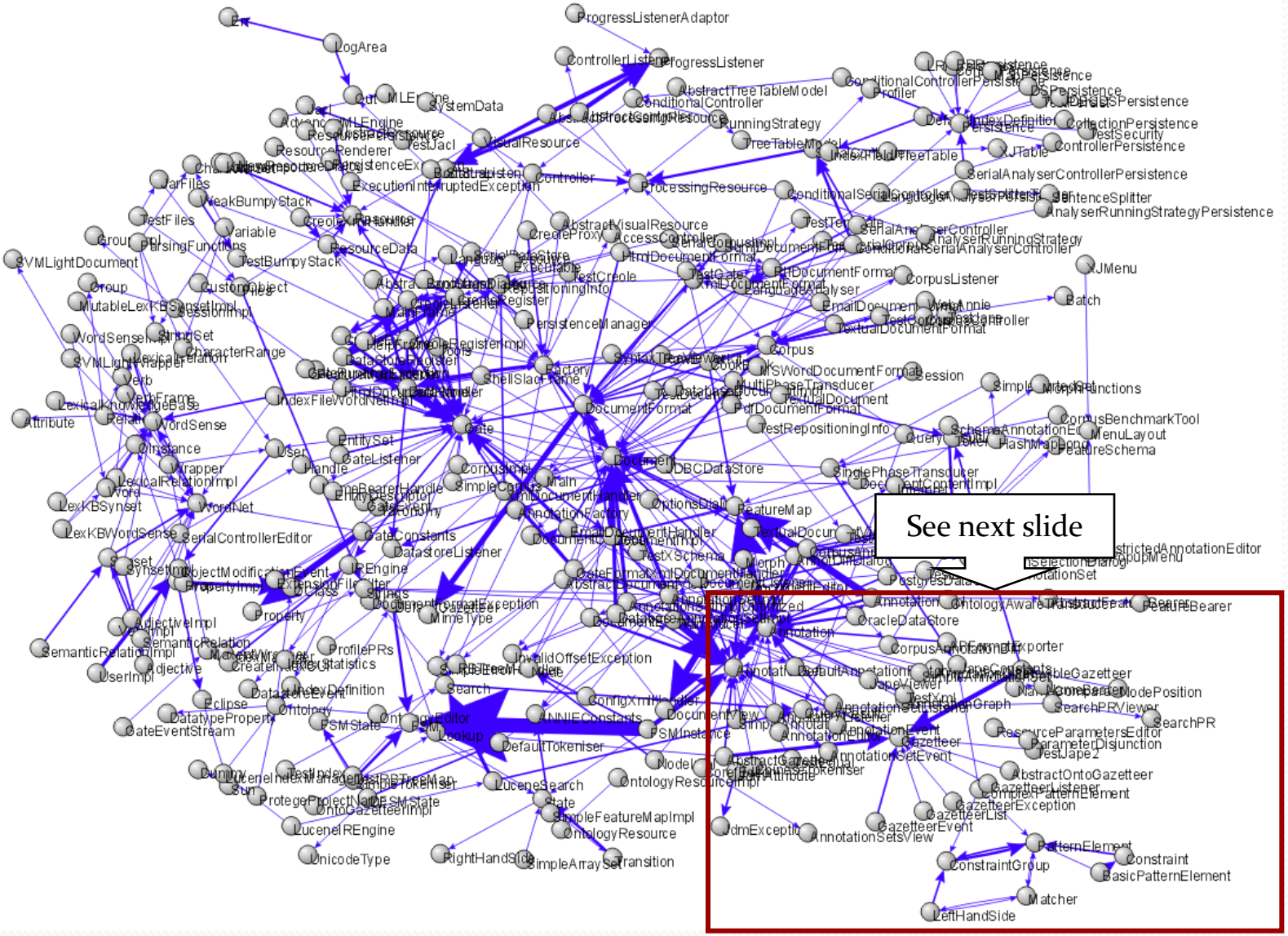
```
/** The MIME type of this format. */
 private MimeType mimeType = null;
```

```
/**
   * Find a DocumentFormat implementation that deals with a particular
   * MIME type, given that type.
   * @param  aGateDocument this document will receive as a feature
   *                       the associated Mime Type. The name of the feature is
   *                       MimeType and its value is in the format type/subtype
   * @param  mimeType the mime type that is given as input
   */
 static public DocumentFormat getDocumentFormat(gate.Document aGateDocument,
                                                MimeType mimeType){
```

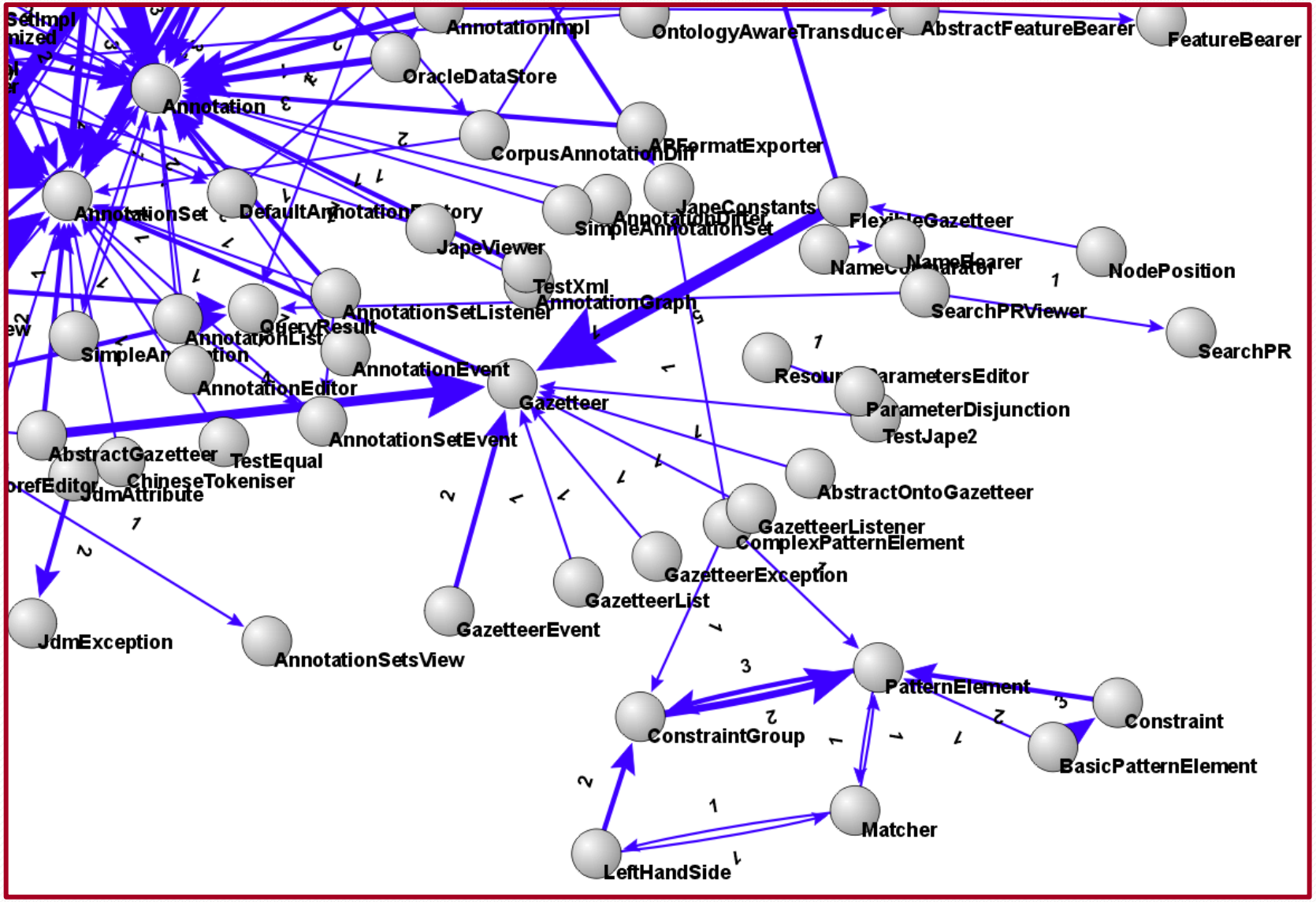
```
 } // getDocumentFormat(aGateDocument, MimeType)

} // class DocumentFormat
```



GATE Comment Reference Network



GATE Comment Reference Network



OntoGen -- Text Garden

File About

Concepts

New Move Delete

- root
  - annotation, event, document
  - exception, impl, words
  - test, indexed, property

Concept properties

Details | Suggestions | Relations

Suggest k-Means | Query | Add

No. suggestions: 3 Docs: All Unused

Keywords	No. ...	[%]

Ontology details

Ontology visualization | Concept's documents | Concept Visualization

Map properties | Zoom mode | Select mode

Data storage

Annotation

GUI

WordNet

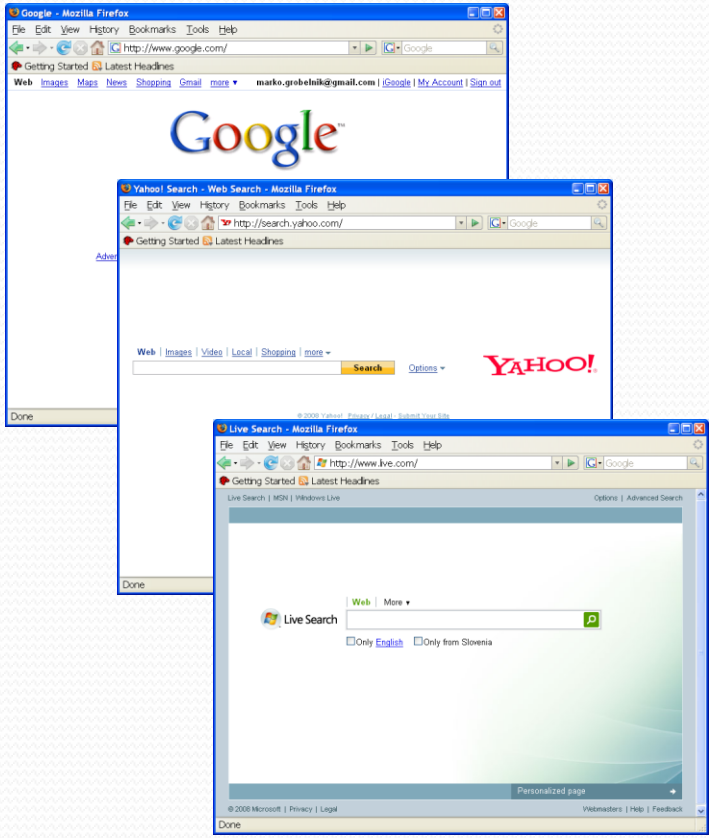
Exceptions

Test

OntoGen news: OntoGen won Best Demo Award at ESWC 2006

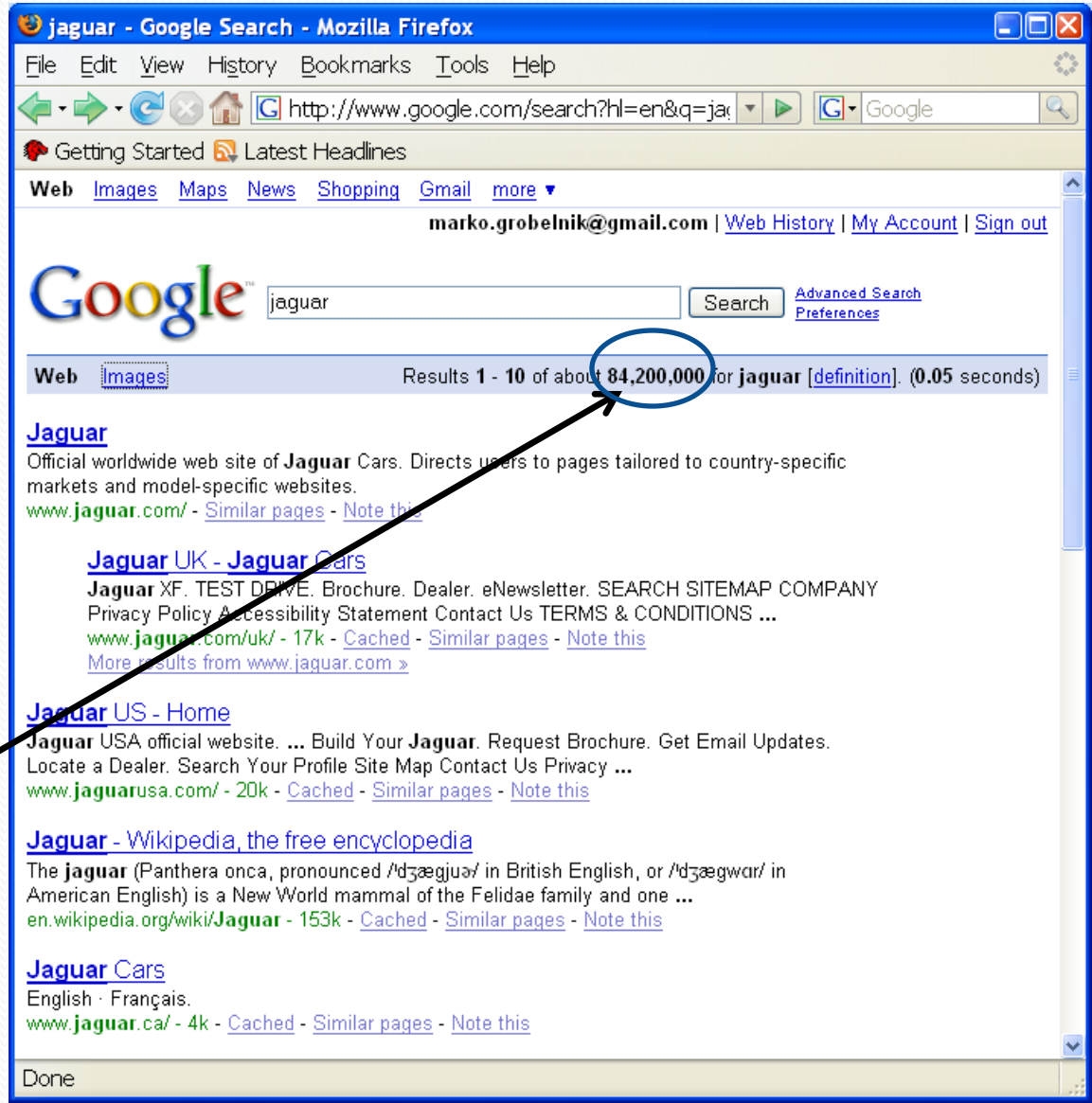
Example: contextualized search

- What are the most common tasks where we manipulate texts in everyday life?
  - “Internet search”!
- …but – how smart is search technology today?
  - …not too smart!
  - It is sophisticated, but not smart



Example: searching for "Jaguar"

- Query "**jaguar**" has many meanings…
- …but the first page of search engines doesn't provide us with many answers
- …there are 84M more results



Context sensitive search with http://searchpoint.ijs.si

Query

Conceptual map

Search Point

Dynamic contextual ranking based on the search point

jaguar - SearchPoint - Windows Internet Explorer

http://searchpoint.ijs.si/Result.aspx

jaguar

Search via topics | Search via query to ontology | Search via hits to ontology

(9) Jaguar
General information and facts from Big Cats Online.
http://www.abf90.dial.pipex.com/jaguar.htm

(59) Jaguar, Jaguar Profile, Facts, Information, Photos, Pictures ...
Get jaguar profile, facts, information, photos, pictures, sounds, habitats, reports, news, and more from National Geographic.
http://animals.nationalgeographic.com/animals/mammals/jaguar.html

(6) Jaguar - Wikipedia, the free encyclopedia
The jaguar (Panthera onca) is a New World mammal of the Felidae family and one of four "big cats" in the Panthera genus, along with the tiger, ...
http://en.wikipedia.org/wiki/Jaguar

(11) Jaguar
Jaguar Facts, Jaguar Photos and Jaguars in the news at the world's largest big cat rescue and sanctuary.
http://www.bigcatrescue.org/jaguar.htm

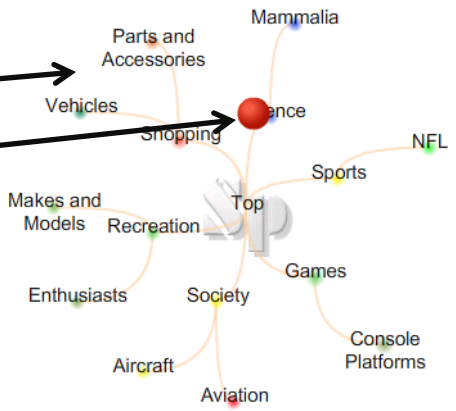
(1) Jaguar
Official worldwide web site of Jaguar Cars. Directs users to pages tailored to country-specific markets.
http://www.jaguar.com/

(32) Jaguar
Contains extensive information about the Jaguar. Information includes habitat, body size, and life span.
http://www.abf90.dial.pipex.com/bco/jaguar.htm

(2) Jaguar UK - Jaguar Cars
Jaguar & Ownership. Highlights. Gallery. Models & Pricing. Design Your XK. TEST DRIVE. Brochure. Dealer. eNewsletter ...
http://www.jaguar.co.uk/

(17) Jaguar Enthusiasts' Club
World's largest audited membership. UK-based, JEC's site has extensive resources available for the enthusiast, including information about their Sections, ...
http://www.jec.org.uk/

(20) San Diego Zoo's Animal Bytes: Jaguar
Get fun and interesting jaguar facts in an easy-to-read style from the San Diego Zoo's Animal



Example: Detecting News Reporting Bias

- The task:
  - Given a news story, are we able to say from which news source it came?

- We compared **CNN** and **Aljazeera** reports about the same events from the war in Iraq
  - …300 aligned articles describing the same story from both sources
- The same topics are expressed in both sources with the following keywords:
  - CNN with:
    - **Insurgents**, Troops, Baghdad, Iran, **Militant**, Police, **Suicide**, **Terrorist**, United, National, Hussein, **Alleged**, Israeli, Syria, Terrorism…
  - Aljazeera with:
    - Attacks, Claims, **Rebels**, Withdrawing, Report, **Fighters**, President, **Resistance**, Occupation, Injured, Army, Demanded, Hit, Muslim, …

Semantics, Knowledge and Common sense reasoning

Towards text understanding…

- The key element to understand the text is to go beyond characters and words…
  - …meaning, we need to have knowledge in the form of a "**world model**" where all the facts from text fit,
  - …we need to be able to deal with **contexts**, and
  - …we need to be able to **reason**

- Do we have something which would go in this direction?
  - …there were couple of trials in the last decades
  - …the only marketable system is **Cyc** from a company **CyCorp** (US and Europe/Slovenia based)
    - New York Times article on Cyc and Web 3.0:
      - http://www.nytimes.com/2006/11/12/business/12web.html

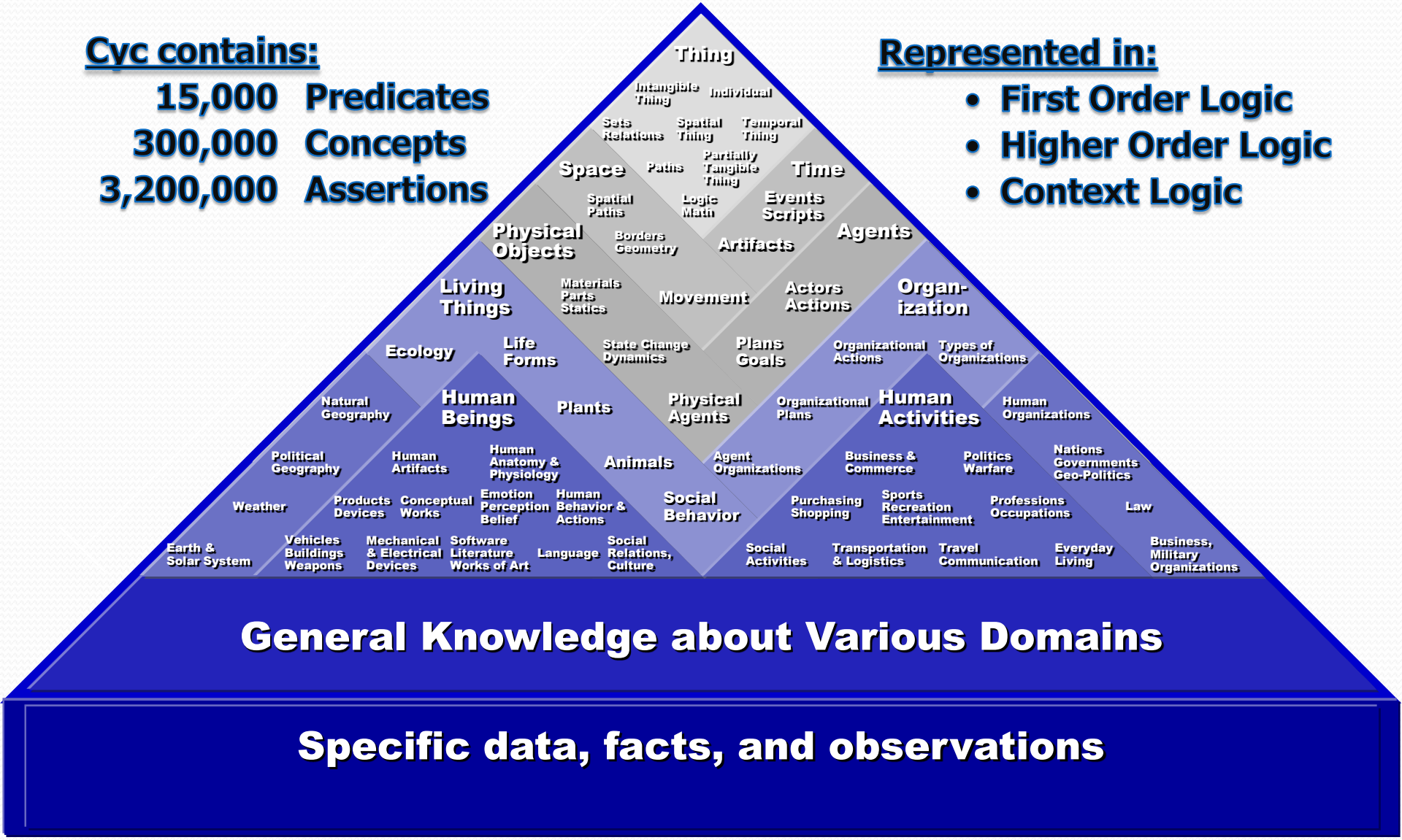
The Cyc Ontology – knowledge about common sense

Cyc contains:

- 15,000 Predicates
- 300,000 Concepts
- 3,200,000 Assertions

Represented in:

- First Order Logic
- Higher Order Logic
- Context Logic



Cyc's front-end: "**Cyc Analytic Environment**" – querying (1/2)

Cyc Analytical Environment: General Intelligence Analysis using the CAE

File Edit Tools Window Help

Task Info | Document Search | Concepts | Related-to Query Creator | Queries

Find | Stop

- Search Results
- Cyc Query Library
  - Overviews
  - Pre-formed Examples
    - Example Hezbollah Queries
    - Actions by Specific Terrorist Groups
    - Support for/by Terrorist Groups
    - Terrorist Group Membership/Leadership
    - Terrorist Acts in a Specific Region
    - Terrorist Group Areas of Operation
    - Terrorist Group Tactics
    - Terrorist Group Suborganizations
    - Casualties in Terrorist Attacks
    - Affiliations with Terrorist Agents
    - Terrorist Group Ideologies
    - Terrorist Attack Targets
    - Weapons Used in Terrorist Attacks
      - List all bombings after 2000 and before September 2001 that used pipe bombs.
    - Motives for Terrorist Attacks
      - Who has a motive for the assassination of Rafik Hariri?
    - Locations of Terrorist Attacks
      - What suicide bombings occurred in what cities in 2004?
    - Responsibility for Terrorist Attacks
      - List the known bombings in which the performer has claimed responsibility.
  - Statistical
    - What percentage of bombings in Northern Ireland were committed by the IRA?
    - Between March and April 2002, in Colombia, what percentage of kidnapping attacks directed at public officials were perpetrated by FARC?
    - For each major attack type, what is the ratio of Hamas attacks that are of that type?
    - What percentage of kidnappings in Israel were perpetrated by Hamas?
    - What percentage of Al Qaida bombings are suicide bombings?
    - List the ratio of suicide bombings to regular bombings by terrorist groups that operate in Afghanistan.
    - List the Islamic Jihad organization with the highest total wounded in their attacks.
    - Which group perpetrated the largest hostage-taking in Europe?
    - List the terrorist group that operates out of Pakistan that has the highest casualty count.
    - What was the deadliest suicide attack?
    - What is the shortest length of time between events performed by suborganizations of Hezbollah?
    - What is the longest duration of time between events performed by suborganizations of Hezbollah?
    - What Islamic Jihad Organization killed the most people?
    - What is the ratio of Hamas attacks that target buildings that are performed in Israel?
    - What is the number of people killed in terrorist attacks in Syria?
    - What percentage of the suicide attacks in Israel are performed by Hamas?
    - Give the total number of people wounded in attacks by the terrorist group People Against Gangsterism And Drugs.
    - What is the number of suicide bombings that occur in Beirut?
  - January 2006 Analyst Session
  - Links between entities
  - AKB-SME research
  - General Purpose

Text query

***WHO* had a motive for the assassination of Hariri.**

Continue | Save | New Tab | Reset

5 answers
Timed out

☑ Allow speculation?

Query (semi) automatically translated in the First Order Logic

Answers (5)

Answer	Speculation Level	Sources
Bashar al-Assad	No Speculation	W CNN
Syria	Mildly Speculative	CNN
al Qaeda	Moderately Speculative	SAIC CNN
United States, the	No Speculation	2
Israel	No Speculation	2

Justify | Fact Sheet | Visualize | Visualize All

Answers to the query

Status: Finished Message: No appropriate visualizations found

Cyc's front-end: "**Cyc Analytic Environment**" – justification (2/2)

Cyc Analytical Environment: General Intelligence Analysis using the CAE

File Edit Tools Window Help

Task Info | Document Search | Concepts | Related-to Query Creator | Queries | Justification | Justification | Justification

Proof 1 | Save... | Copy

Query: Who or what had a motive for the assassination of Hariri?
Answer: al Qaeda

Because:

Since 2000, Lebanon has been responsible for according with Lebanese economic reform. 1

February 14, 2005 was the date of the assassination of Hariri. 2

Rafik Hariri was killed during the assassination of Hariri. 2
Rafik Hariri is an advocate of Lebanese economic reform.

Al Qaeda opposes Lebanese economic reform.

▼ Options

▼ Options

Detailed Justification:

▶ Al Qaeda had a motive for the assassination of Hariri.

External Sources:

1 Gary C. Gambill, "Dossier: Rafiq Hariri", *United States Committee for a Free Lebanon,* July 2001, http://www.meib.org/articles/0107_id1.htm .

2 "Huge blast kills Lebanese ex-PM", *the Cable News Network,* February 14, 2005, http://www.cnn.com/2005/WORLD/meast/02/14/beirut.explosion.1910/ .

Query & Answer

Justification

Sources for Reasoning and Justification

Further online information

Recorded tutorials, lectures, summer-schools available from http://videolectures.net

- Semantic Web: http://videolectures.net/Top/Computer_Science/Semantic_Web/