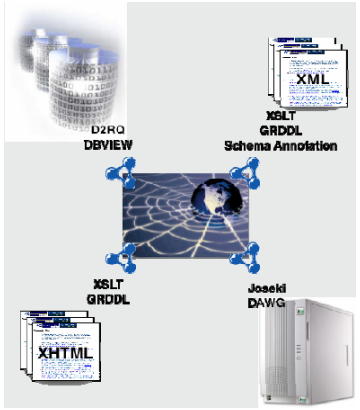
MỘT SỐ HƯỚNG NGHIÊN CỨU VÀ ỨNG DỤNG

Web ngữ nghĩa

- Mục tiêu: phát triển các chuẩn chung và công nghệ cho phép máy tính có thể hiểu được nhiều hơn thông tin trên Web, sao cho chúng có thể hỗ trợ tốt hơn việc khám phá thông tin, tích hợp dữ liệu, và tự động hóa các công việc.



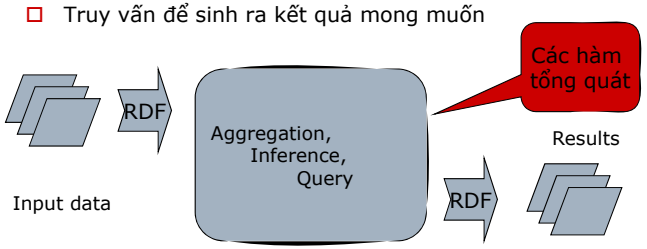
Các loại ứng dụng

- Các dạng dữ liệu bán cấu trúc
- Các ứng dụng mở: thêm các chức năng mới với các loại dữ liệu cũ và mới
- Ví dụ:
  - Quản lý thông tin cá nhân (Chandler)
  - Mạng xã hội (FOAF)
  - Tổ chức thông tin (RSS,PRISM)
  - Dữ liệu thư viện/bảo tàng (Dublin Core, Harmony)

Những gì có thể làm được

Nếu dữ liệu đầu vào ở dạng RDF, các hàm sau có thể thực hiện

- Tích hợp nhiều nguồn dữ liệu
- Suy diễn để sinh ra thông tin mới
- Truy vấn để sinh ra kết quả mong muốn



Aggregation + Inference = New Knowledge

- Building on the success of XML
  - Common syntactic framework for data representation, supporting use of common tools
  - But, lacking semantics, provides no basis for automatic aggregation of diverse sources
- RDF: a semantic framework
  - Automatic aggregation (graph merging)
  - Inference from aggregated data sources generates new knowledge
    - Domain knowledge from ontologies and inference rules

5

Aggregation + Inference: Example

- Consider three datasets, describing:
  - vehicles' passenger capacities
  - the capacity of some roads
  - the effect of policy options on vehicle usage
- Aggregation and inference may yield:
  - passenger transportation capacity of a given road in response to various policy options
  - using existing open software building blocks

6

What needs to be done?

- Information design
- Data-use strategies and inference rules
- Mechanisms for acquisition of existing data sources
- Mechanisms for presentation or utilization of the resulting information

7

Benefits

- Greater use of off-the-shelf software
  - reduced development cost and risk
- Re-use of information designs
  - reduced application design costs; better information sharing between applications
- Flexibility
  - systems can adapt as requirements evolve
- Open access to information making possible new applications

8

Recommendation: Low risk approach

- Focus on information requirements
  - this is unlikely to be wasted effort
- Start with a limited goal, progress by steps
  - adapting to evolving requirements is an advantage of SW technology; if it can do this for large projects it certainly must be able to do so for early experimental projects
- Use existing open building blocks

Lots of Tools (*not an exhaustive list!)*

Categories:

- Triple Stores
- Inference engines
- Converters
- Search engines
- Middleware
- CMS
- Semantic Web browsers
- Development environments
- Semantic Wikis
- …

Some names:

- Jena, AllegroGraph, Mulgara,
- Sesame, flickurl, …
- TopBraid Suite, Virtuoso
- environment, Falcon, Drupal 7,
- Redland, Pellet, …
- Disco, Oracle 11g, RacerPro,
- IODT, Ontobroker, OWLIM, Talis
- Platform, …
- RDF Gateway, RDFLib, Open
- Anzo, DartGrid, Zitgist, Ontotext,
- Protégé, …
- Thetus publisher, SemanticWorks,
- SWI-Prolog, RDFStore…
- …

Application patterns

- It is fairly difficult to "categorize" applications
- Some of the application patterns:
  - data integration
  - intelligent (specialized) Web sites (portals) with improved local search
  - content and knowledge organization
  - knowledge representation, decision support
  - data registries, repositories
  - collaboration tools (eg, social network applications)

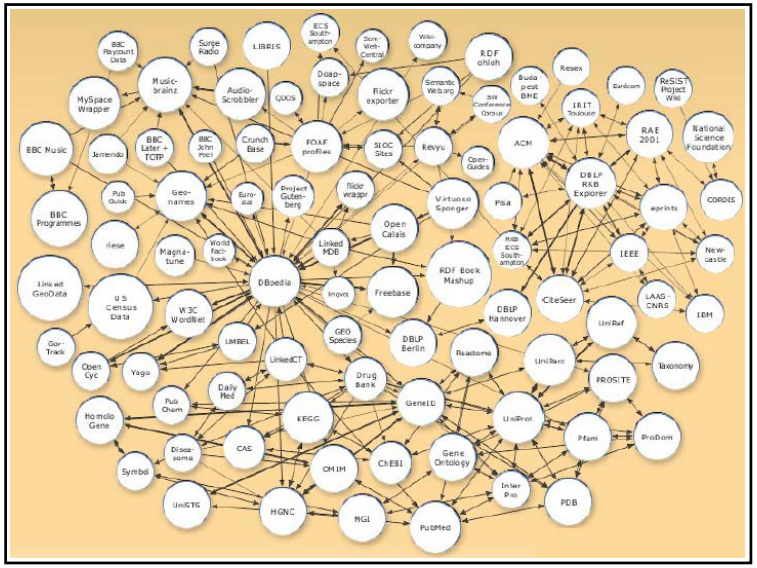
To "seed" a Web of Data...

- Data has to be published, ready for integration
- And this is now happening!
  - Linked Open Data project
  - eGovernmental initiatives in, eg, UK, USA, France,...
  - Various institutions publishing their data

Linking Open Data Project

- Goal: "expose" open datasets in RDF
- Set RDF links among the data items from different datasets
- Set up SPARQL Endpoints
- Billions triples, millions of "links"

13



Example data source: DBpedia

- DBpedia is a community effort to extract structured ("infobox") information from Wikipedia
- provide a SPARQL endpoint to the dataset
- interlink the DBpedia dataset with other datasets on the Web

15

Extracting structured data from Wikipedia


```
@prefix dbpedia <http://dbpedia.org/resource/>.
@prefix dbterm  <http://dbpedia.org/property/>.

dbpedia:Amsterdam
   dbterm:officialName "Amsterdam" ;
   dbterm:longd "4" ;
   ...
   dbterm:leaderName dbpedia:Job_Cohen ;
   ...
   dbterm:areaTotalKm "219" ;
   ...
dbpedia:ABN_AMRO
   dbterm:location dbpedia:Amsterdam ;
   ...
```

16

Automatic links among open datasets

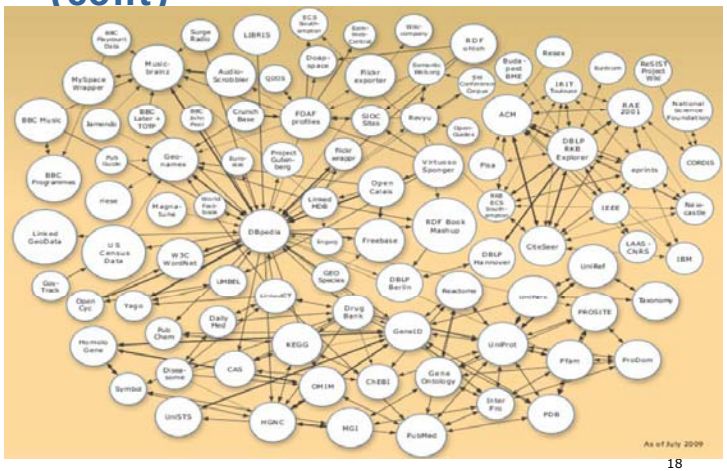
```
<http://dbpedia.org/resource/Amsterdam>
  owl:sameAs <http://rdf.freebase.com/ns/...> ;
  owl:sameAs <http://sws.geonames.org/2759793> ;
  ...

<http://sws.geonames.org/2759793>
  owl:sameAs <http://dbpedia.org/resource/Amsterdam>
  wgs84_pos:lat "52.3666667" ;
  wgs84_pos:long "4.8833333" ;
  geo:inCountry <http://www.geonames.org/countries/#NL> ;
...
```

Processors can switch automatically from one to the other…

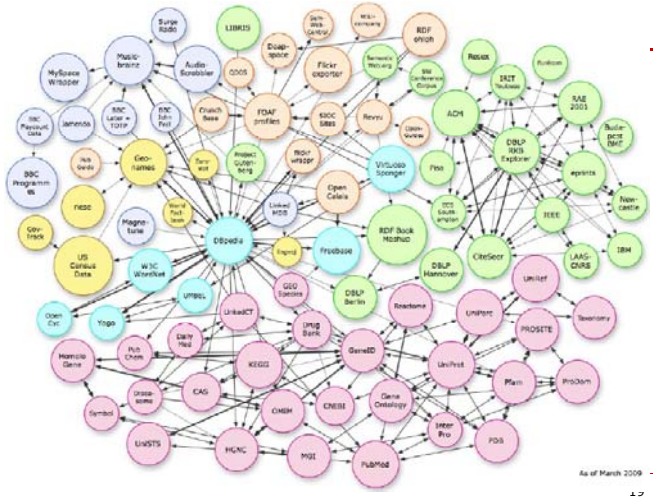
17

Linking Open Data Project (cont)



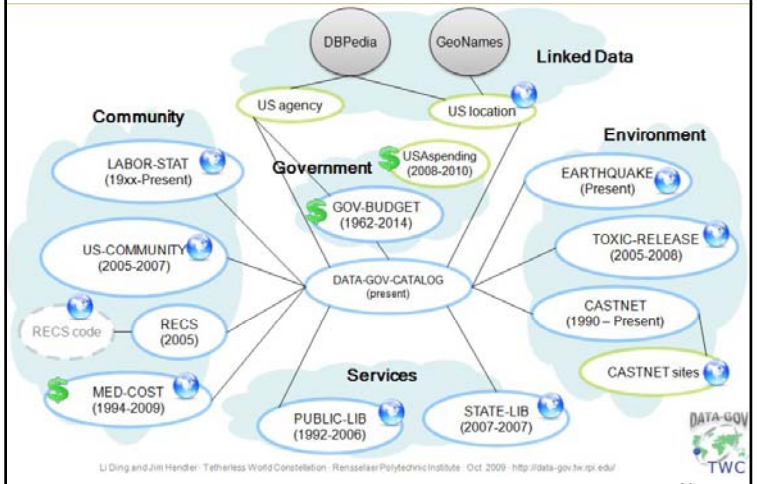
18

Linking Open Data Project (cont)



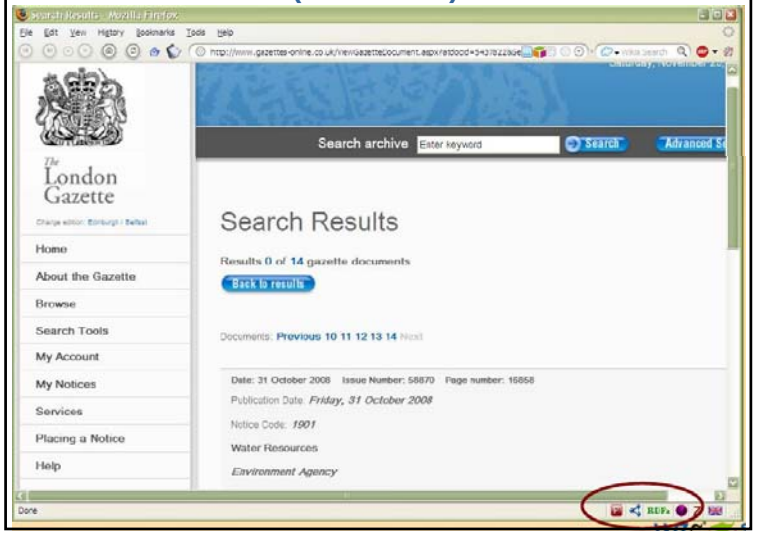
19

Linked Open eGov Data

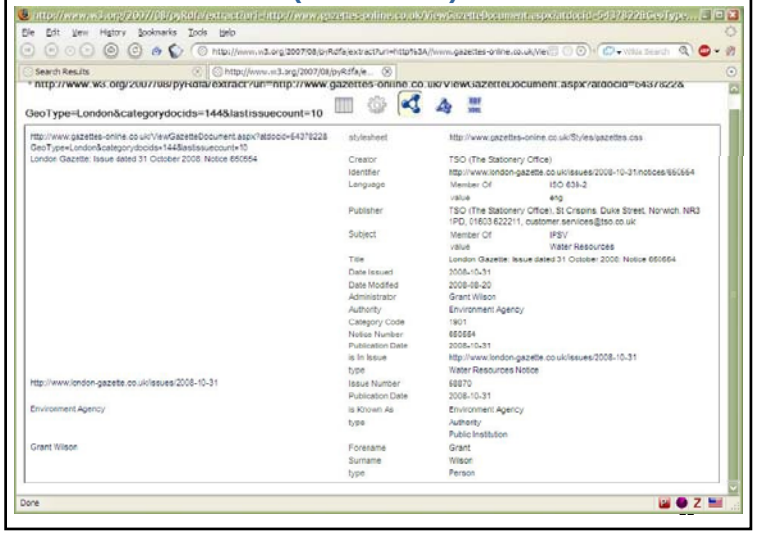


20

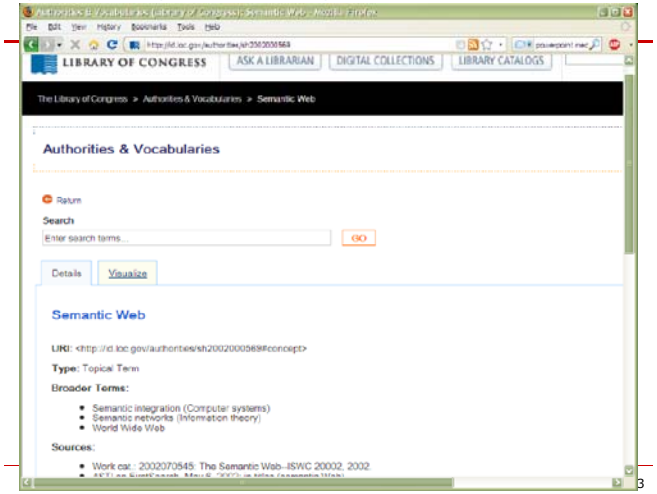
Publication of data (with RDFa): London Gazette



Publication of data (with RDFa): London Gazette

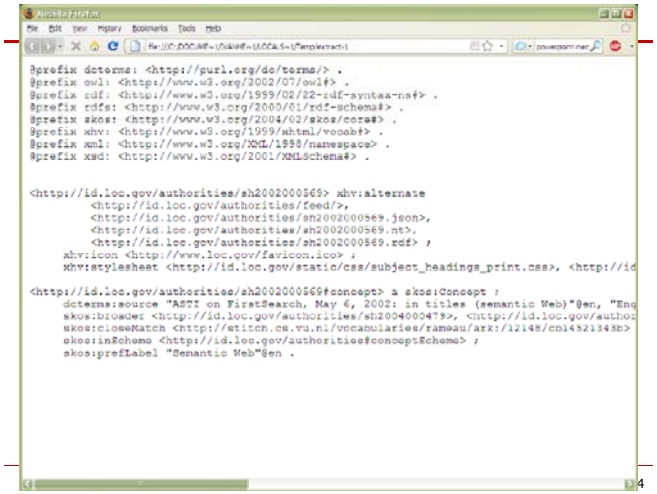


Publication of data (with RDFa & SKOS): Library of Congress Subject Headings



3

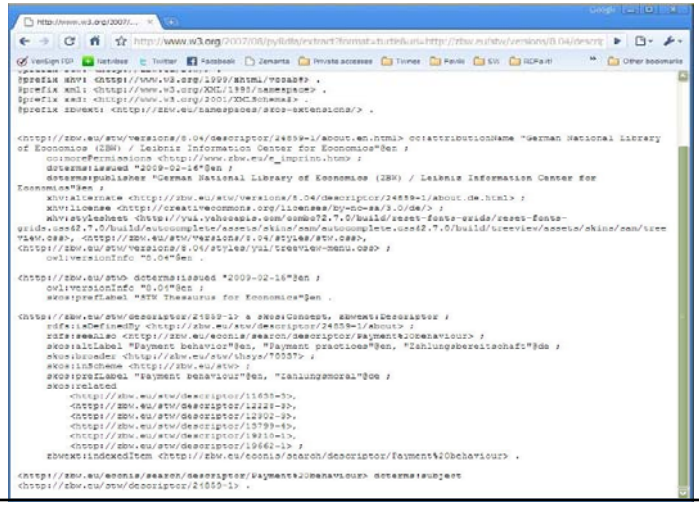
Publication of data (with RDFa & SKOS): Library of Congress Subject Headings



4

Publication of data (with RDFa & SKOS):Economics Thesaurus

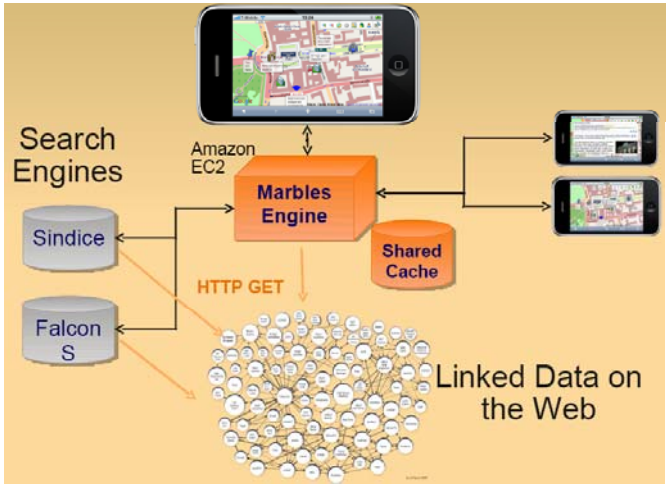
Publication of data (with RDFa & SKOS):Economics Thesaurus



Using the LOD cloud on an iPhone

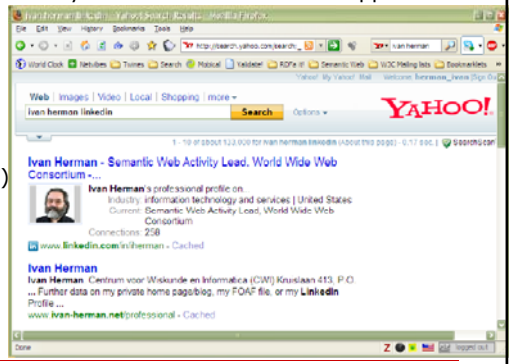
Using the LOD cloud on an iPhone

Using the LOD cloud on an iPhone



You publish the raw data, W3C use it…

- **Yahoo's SearchMonkey**
- Search based results may be customized via small applications
- Metadata embedded in pages (in RDFa, eRDF, etc) are reused
- Publishers can export extra (RDF) data via other formats



30

Google's rich sniplet

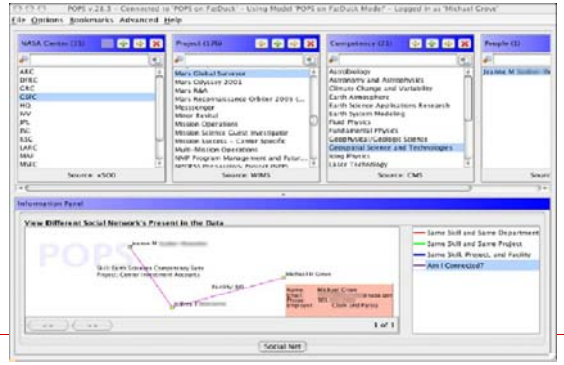
- Embedded metadata (in microformat or RDFa) is used to improve search result page
  - at the moment only a few vocabularies are recognized, but that will evolve over the years

Drooling Dog Bar B Q - Colfax, CA
★★★★☆ 15 reviews - Price range: $$
Drooling Dog has some really good **BBQ**. I had the pulled pork sandwich, ... **Drooling Dog BBQ** is a great place to stop at on your way up the hill to Tahoe ...
www.yelp.com/biz/**drooling-dog**-bar-b-q-colfax - 75k - Cached - Similar pages

31

Find experts at NASA

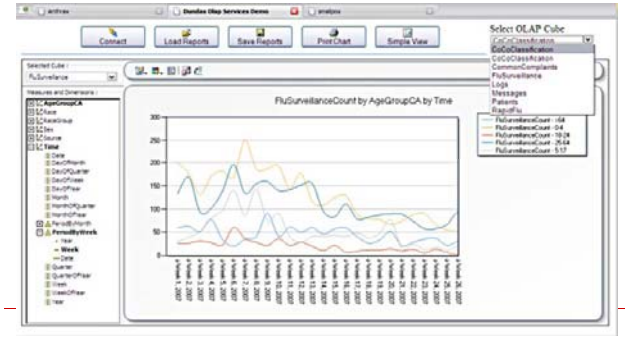
- Expertise locater for nearly 70,000 NASA civil servants
- over 6 or 7 geographically distributed databases, data sources, and web services…



32

Public health surveillance (Sapphire)

- Integrated biosurveillance system (biohazards, bioterrorism, disease control, etc)
- Integrates multiple data sources
- new data can be added easily



33

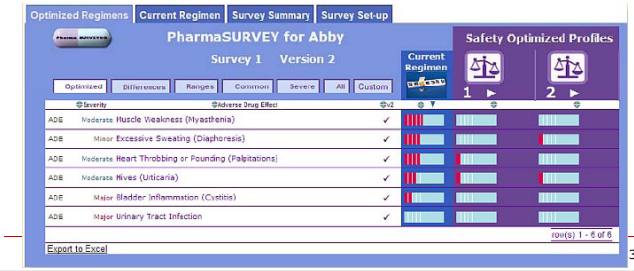
A frequent paradigm: intelligent portals

- "Portals" collecting data and presenting them to users
- They can be public or behind corporate firewalls
- Portal's internal organization makes use of semantic data, ontologies
  - integration with external and internal data
  - better queries, often based on controlled vocabularies or ontologies…

34

Help in choosing the right drug regimen

- Help in finding the best drug regimen for a specific case, per patient
- Integrate data from various sources (patients, physicians, Pharma, researchers, ontologies, etc)
- Data (eg, regulation, drugs) change often, but the tool is much more resistant against change

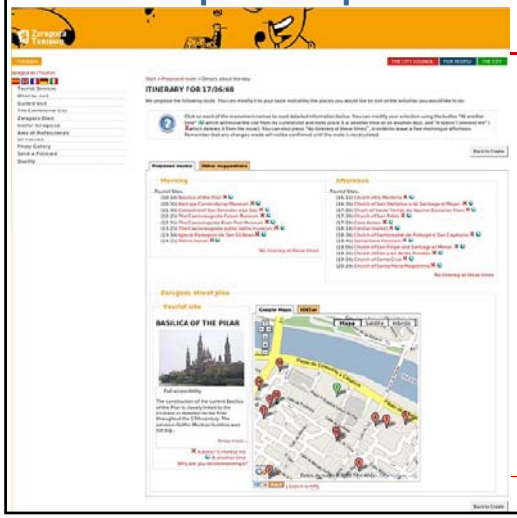


35

Portal to aquatic resources

36

eTourism: provide personalized itinerary

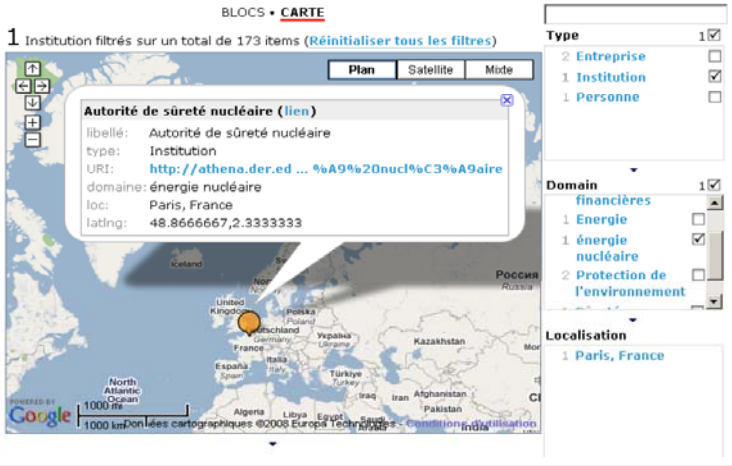


- Integration of relevant data in Zaragoza (using RDF and ontologies)
- Use rules on the RDF data to provide a proper itinerary

Integration of "social" software data

- Internal usage of wikis, blogs, RSS, etc, at EDF
- goal is to manage the flow of information better
- Items are integrated via
  - RDF as a unifying format
  - simple vocabularies like SIOC, FOAF, MOAT (all public)
  - internal data is combined with linked open data like Geonames
  - SPARQL is used for internal queries
- Details are hidden from end users (via plugins, extra layers, etc)

Integration of "social" software data



Improved Search via Ontology (GoPubMed)

- Search results are re-ranked using ontologies
- Related terms are highlighted, usable for further search

New type of Web 2.0 applications

- New Web 2.0 applications come every day
- Some begin to look at Semantic Web as possible technology to improve their operation
  - more structured tagging, making use of external services
  - providing extra information to users
  - etc.
- Some examples: Twine, Revyu, Faviki, …

"Review Anything"

Faviki: social bookmarking, semantic tagging

- Social bookmarking system (a bit like del.icio.us) but with a controlled set of tags
  - tags are terms extracted from wikipedia/Dbpedia
  - tags are categorized using the relationships stored in Dbpedia
  - tags can be multilingual, DBpedia providing the linguistic bridge
- The tagging process itself is done via a user interface hiding the complexities

Other application areas come to the fore

- Content management
- Business intelligence
- Collaborative user interfaces
- Sensor-based services
- Linking virtual communities
- Grid infrastructure
- Multimedia data management
- Etc

CEO guide for SW: the "DO-s"

- **Start small:** Test the Semantic Web waters with a pilot project […] before investing large sums of time and money.
- **Check credentials:** A lot of systems integrators don't really have the skills to deal with Semantic Web technologies. Get someone who's savy in semantics.
- **Expect training challenges:** It often takes people a while to understand the technology. […]
- **Find an ally:** It can be hard to articulate the potential benefits, so find someone with a problem that can be solved with the Semantic Web and make that person a partner.

CEO guide for SW: the "DON'T-s"

- **Go it alone:** The Semantic Web is complex, and it's best to get help.
- **Forget privacy:** Just because you can gather and correlate data about employees doesn't mean you should. Set usage guidelines to safeguard employee privacy.
- **Expect perfection:** While these technologies will help you find and correlate information more quickly, they're far from perfect. Nothing can help if data are unreliable in the first place.
- **Be impatient:** One early adopter at NASA says that the potential benefits can justify the investments in time, money, and resources, but there must be a multi-year commitment to have any hope of success

Web ngữ nghĩa

- Nghiên cứu về Web ngữ nghĩa:
  - Chuẩn hoá các ngôn ngữ biểu diễn dữ liệu (XML) và siêu dữ liệu (RDF) trên Web.
  - Chuẩn hoá các ngôn ngữ biểu diễn Ontology cho Web có ngữ nghĩa.
  - Phát triển nâng cao Web có ngữ nghĩa (Semantic Web Advanced Development - SWAD).

Web ngữ nghĩa

- SWAD: làm thế nào để nhúng ngữ nghĩa một cách tự động vào các tài liệu Web?
  - trích tự động ngữ nghĩa của mỗi tài liệu Web
  - Chuyển sang các mẫu chung sử dụng ngôn ngữ web ngữ nghĩa
- Việc tìm kiếm hiệu quả hơn.
  - Ví dụ: tìm thành phố Sài Gòn: trả về các tài liệu có TP.HCM hoặc Sài Gòn như một thành phố, chứ không phải các tài liệu chứa từ "Sài Gòn" như trong "Đội bóng Cảng Sài Gòn", "Xí nghiệp may Sài Gòn", hay "Cty Saigon Tourist".

KIM - Knowledge and Information Management

- KIM của Ontotext Lab, Bulgaria
  - Trích rút thông tin từ các tin tức quốc tế
  - Ontology có ~250 lớp, 100 thuộc tính.
  - CSTT có ~ 80,000 thực thể về các nhân vật, thành phố, công ty, và tổ chức
- VN-KIM: trích rút thực thể trong các trang báo điện tử tiếng Việt, bao gồm:
  - CSTT về các nhân vật, tổ chức, núi non, sông ngòi, và địa điểm phổ biến ở Việt Nam.
  - Khối trích rút thông tin tự động
  - Khối tìm kiếm thông tin và các trang Web về các thực thể

49

VN-KIM

- CSTT được xây dựng trên nền của Sesame, mã nguồn mở quản lý tri thức theo RDF
- Các tài liệu Web có chú thích ngữ nghĩa được đánh chỉ mục và quản lý bằng mã nguồn mở Lucene(mã nguồn mở bằng Java, cung cấp các chức năng truy vấn hiệu quả)
- Khối trích rút thông tin tự độngđược phát triển dựa trên GATE
- Tham khảo: http://www.dit.hcmut.edu.vn/~tru/VN-KIM/index.htm

50

Where are we now?

- Semantic Web is new technology
  - about 10 years after the original WWW
- Many applications are experimental
- The goals may be inevitable...
  - Applications working together with users' information, not owning it
  - drawing background knowledge from the Web
  - less dependence on hand-coded bespoke software
    - … but the particular technology is not

51