The offer and promises of the Semantic Web

Gijón, Spain, 2007-02-02

Ivan Herman, W3C

“Who is Viviane Redding?”

- You can query the EU’s Information Society portal database for speeches held by commisioners
- Go (manually!) to *another* page (generated from *another* database)
- Click to get to Redding’s page
- All these steps must be made *manually*, although the information is available in *different* databases for automatic processing…
- … but the databases are not integrated

Data(base) Integration

- Data sources (eg, HTML pages, databases, …) are very different in structure, in content
- Lots of applications require managing *several* data sources
  - *after company mergers*
  - *combination of administrative data for e-Government*
  - *biochemical, genetic, pharmaceutical research*
  - *etc.*
- Most of these data are accessible from the Web (though not necessarily public yet)

What Is Needed?

- (Some) data should be available for machines for further processing
- Data should be possibly combined, merged on a Web scale
- Sometimes, data may describe other data (like the library example, using metadata)…
- … but sometimes the data is to be exchanged by itself, like my calendar or my travel preferences
- Machines may also need to *reason* about that data

A rough structure of *data integration*

1. Map the various data onto an abstract data representation
   - *make the data independent of its internal representation…*
2. Merge the resulting representations
3. Start making queries on the whole!
   - *queries that could not have been done on the individual data sets*

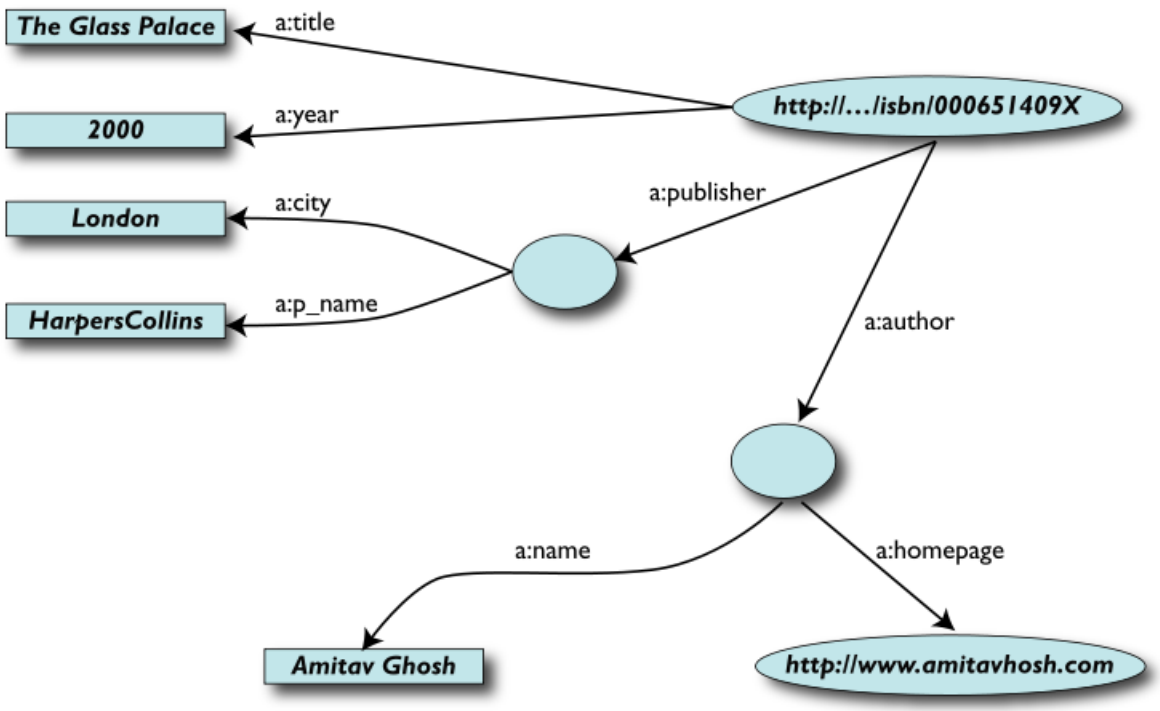
A *simplifed* bookstore data (dataset “A”)

ID	Author	Title	Publisher	Year
ISBN 0-00-651409-X	id_xyz	The Glass Palace	id_qpr	2000

ID	Name	Home page
id_xyz	Amitav Ghosh	http://www.amitavghosh.com/

ID	Publisher Name	City
id_qpr	Harper Collins	London

1st step: export your data as a set of *relations*



Some notes on the exporting the data

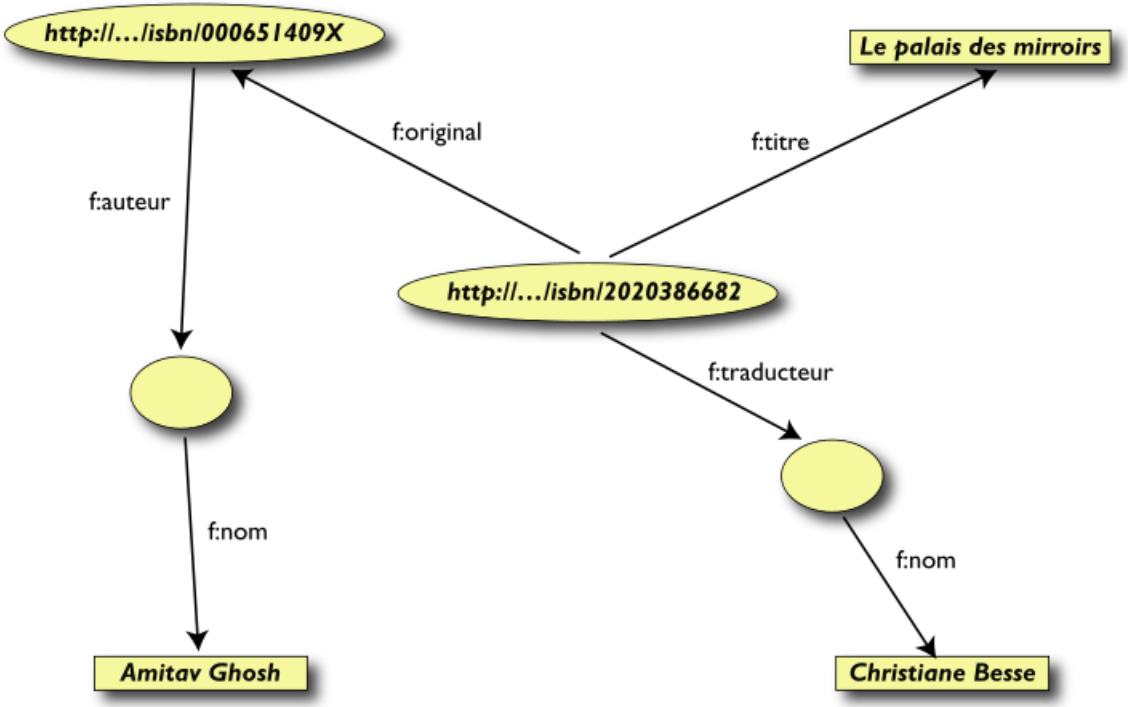
- Relations form a graph
  - *the nodes refer to the "real" data or contain some literal*
  - *how the graph is represented in machine is immaterial for now*
- Data export does *not* necessarily mean physical conversion of the data
  - *relations can be generated on-the-fly at query time*
    - via SQL "bridges"
    - scraping (X)HTML pages
    - extracting data from Excel sheets
    - etc.
- One can export *part* of the data

Another bookstore data (dataset “F”)

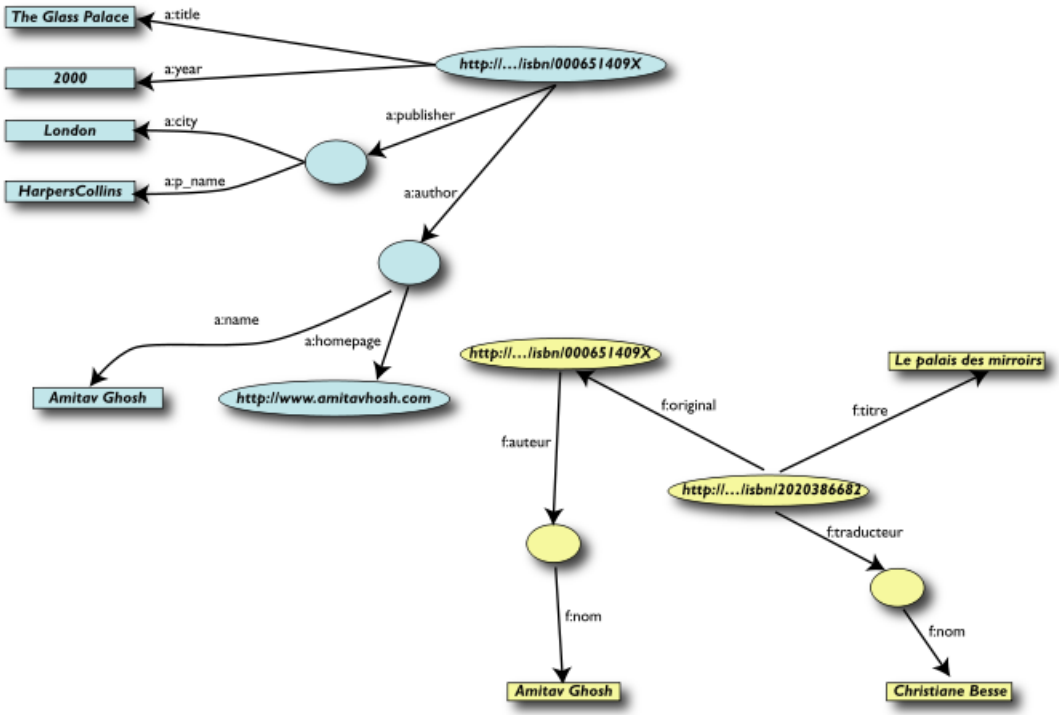
ID	Titre	Auteur	Traducteur	Original
ISBN 2020386682	Le Palais des miroirs	i_abc	i_qrs	ISBN 0-00-651409-X

ID	Nom
i_abc	Amitav Ghosh
i_qrs	Christiane Besse

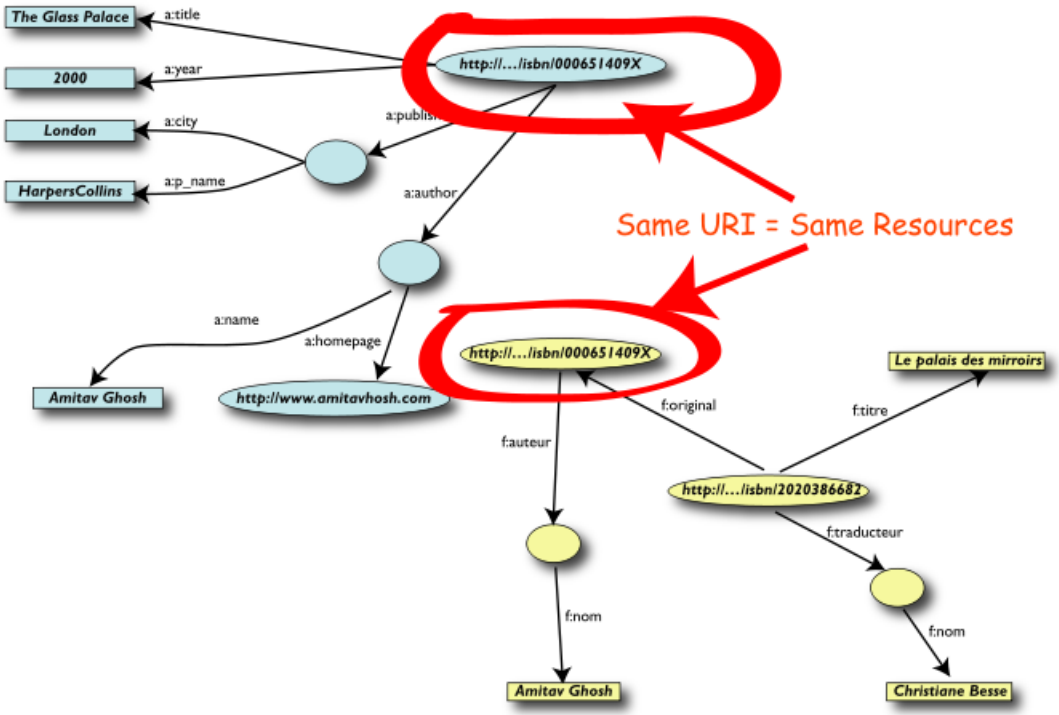
2nd step: export your second set of data



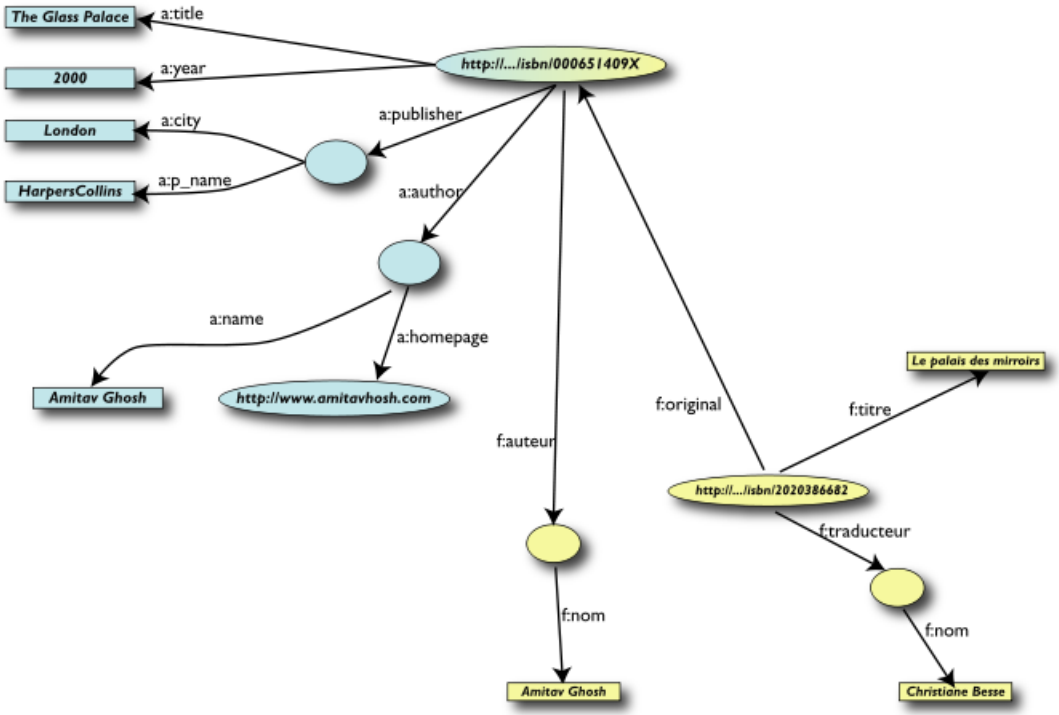
3rd step: start merging your data



3rd step: start merging your data (cont.)

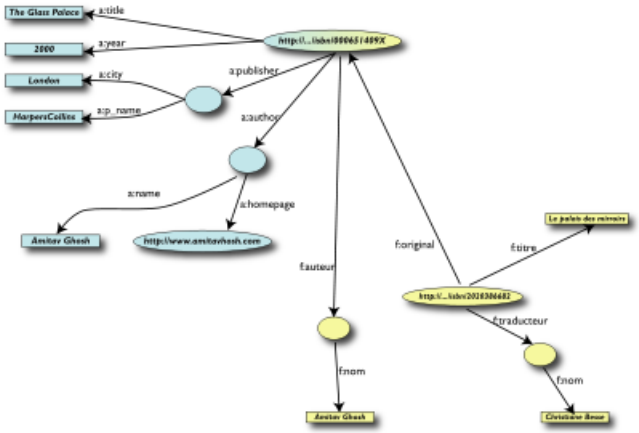


3rd step: merge identical resources



Start making queries…

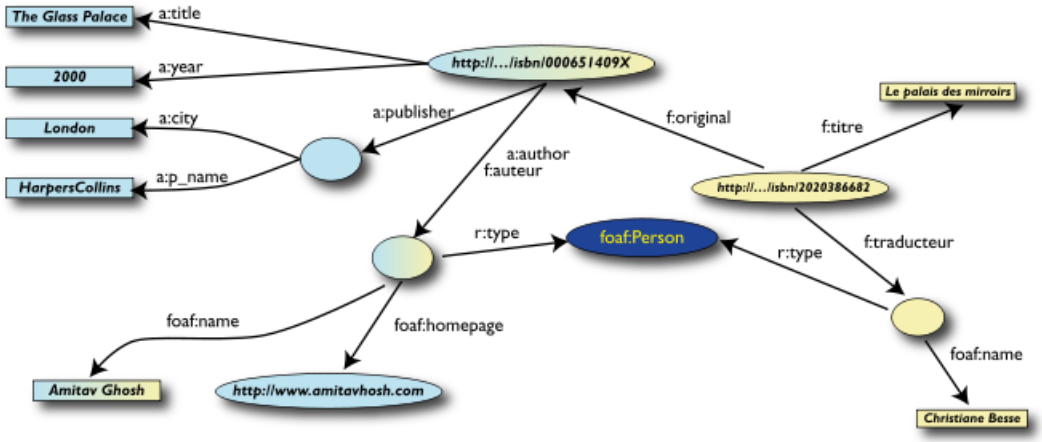
- User of data “F” can now ask queries like:
  - *« donnes-moi le titre de l’original »*
  - *(ie: “give me the title of the original”)*
- This information is not in the dataset “F”…
- …but can be automatically retrieved by merging with dataset “A”!

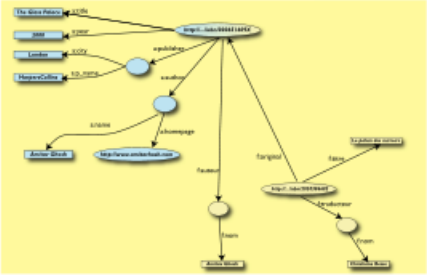


However, more can be achieved…

- We “feel” that `a:author` and `f:auteur` should be the same
- But an automatic merge doest not know that!
- Let us add some extra information to the merged data:
  - `a:author` *same as* `f:auteur`
  - *both identify a “Person”:*
    - a term that a community may have already defined:
      - a “Person” is uniquely identified by his/her name and, say, homepage
      - it can be used as a “category” for certain type of resources

3rd step revisited: use the extra knowledge





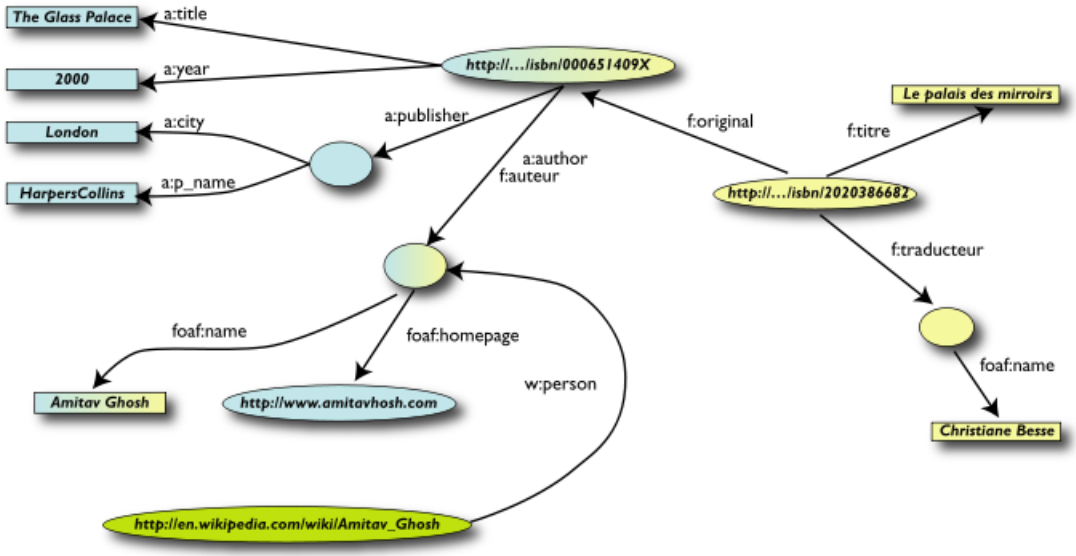
Start making richer queries!

- User of dataset “F” can now query:
  - *« donnes-moi la page d’accueil de l’auteur de l’original »*
  - *(ie, “give me the home page of the original’s author)*
- The data is not in dataset “F”…
- …but was made available by:
  - *merging datasets “A” and datasets “F”*
  - *adding three simple extra statements as an extra “glue”*
  - *using existing terminologies as part of the “glue”*

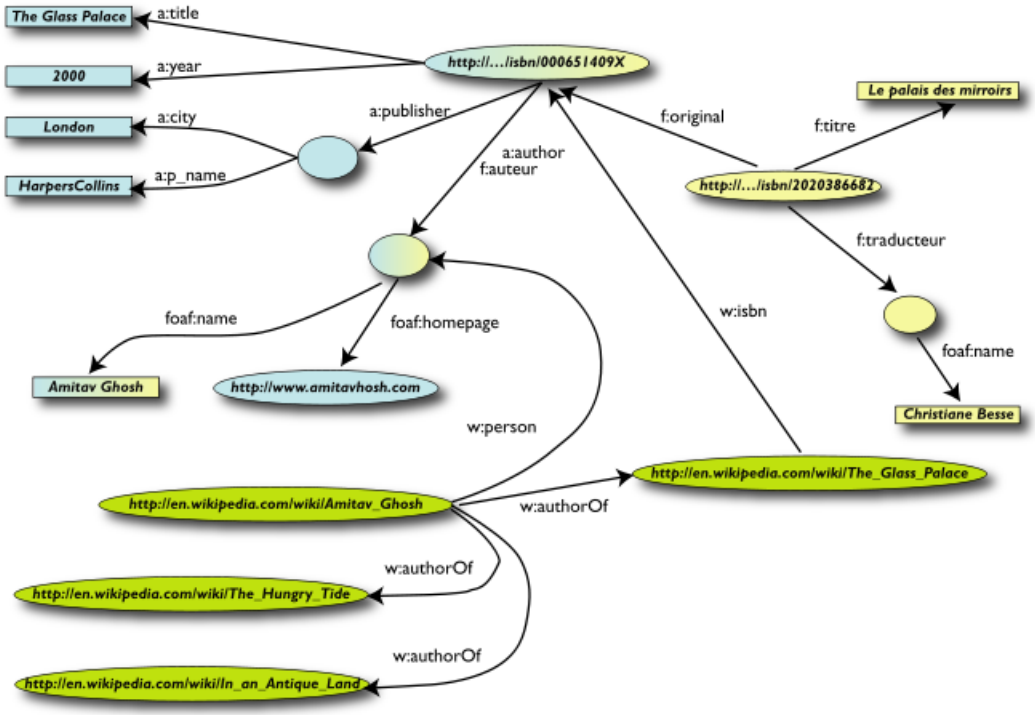
Combine with different datasets

- Using, e.g., the “Person”, the dataset can be combined with other sources
- For example, data in Wikipedia can be extracted using simple (e.g., XSLT) tools
  - *there is an active development to add some simple semantic “tag” to wikipedia entries*
  - *we tacitly presuppose their existence in our example…*

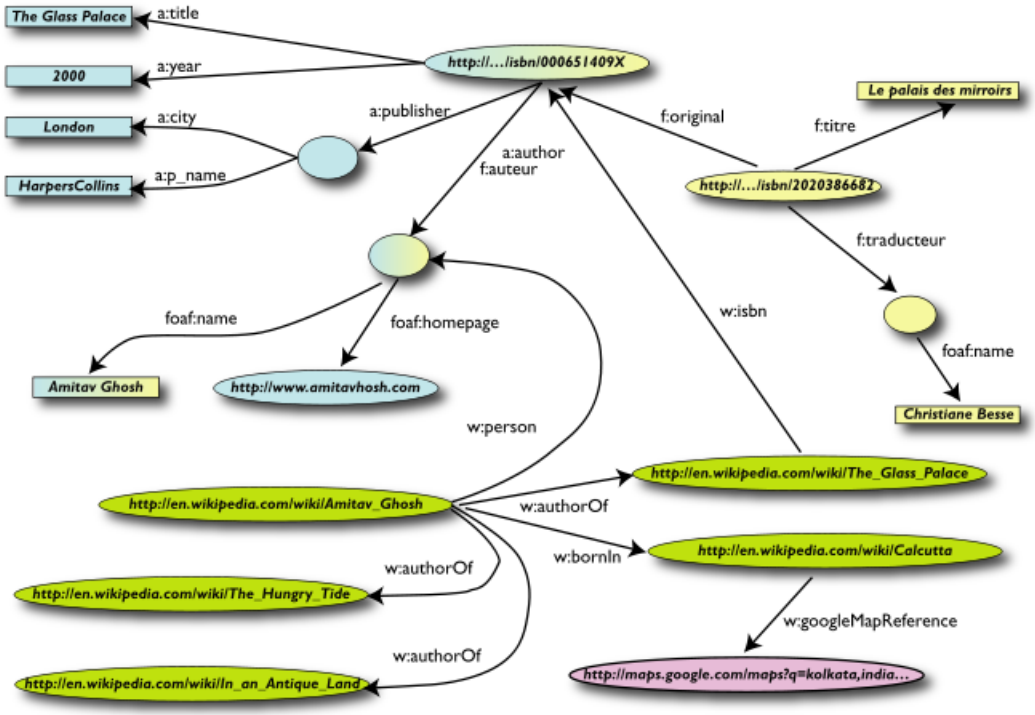
Merge with Wikipedia data



Merge with Wikipedia data



Merge with Wikipedia data



Is that surprising?

- Maybe but, in fact, no…
- What happened via automatic means is done all the time, every day by the users of the Web!
- The difference: a bit of extra rigor (e.g., *naming* the relationships) is necessary so that machines could do this, too

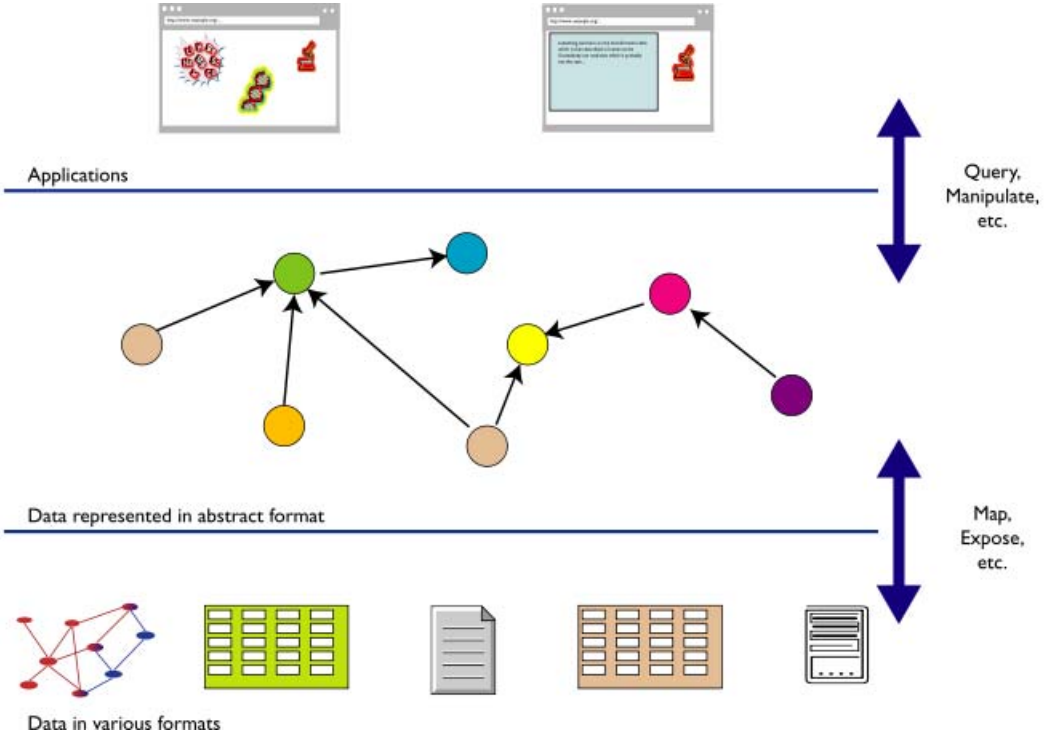
What did we do?

- We combined different datasets
  - *all may be of different origin somewhere on the web*
  - *all may have different formats (mysql, excel sheet, XHTML, etc)*
  - *all may have different names for relations (e.g., multilingual)*
- We could combine the data because some URI-s were identical (the ISBN-s in this case)
- We could add some simple additional information (the “glue”), also using common terminologies that a community has produced
- As a result, *new relations* could be found and retrieved

It could become even more powerful

- We could add extra knowledge to the merged datasets
  - *e.g., a full classification of various type of library data*
  - *geographical information*
  - *etc.*
- This is where *ontologies*, extra *rules*, etc, may come in
- Even more powerful queries can be asked as a result

What did we do? (cont)



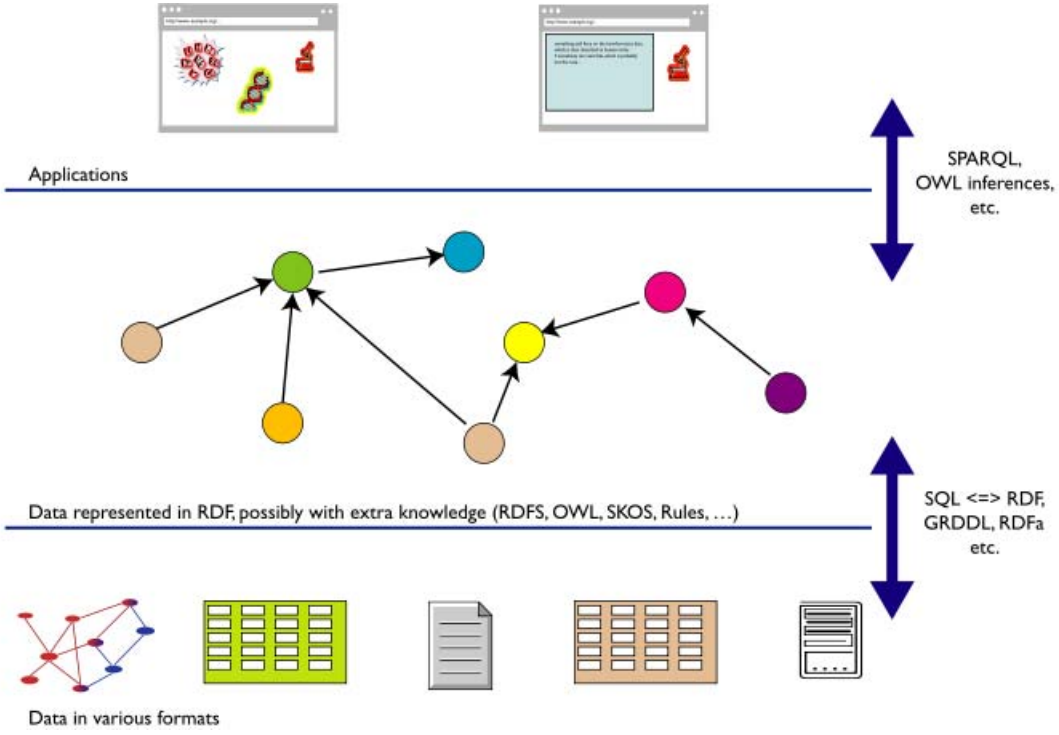
The abstraction pays off because…

- … the graph representation is independent on the *exact* structures in, say, a relational database
- … a change in local database schemas, XHTML structures, etc, do *not* affect the whole, only the "export" step
  - *"schema independence"*
- … new data, new connections can be added seamlessly, regardless of the structure of other datasources

So where is the Semantic Web?

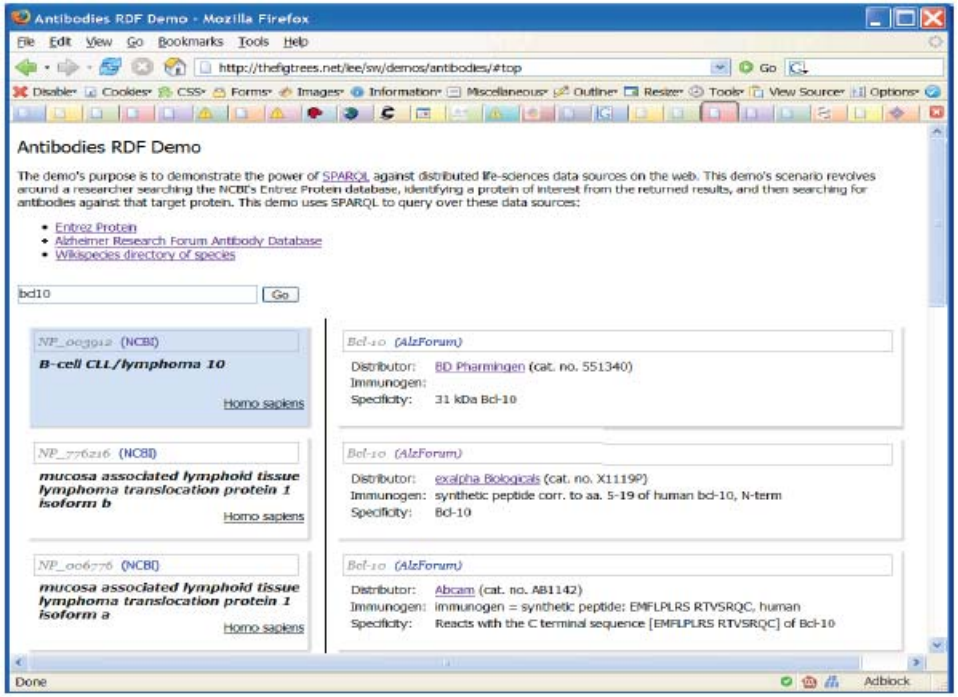
- The Semantic Web provides technologies to make such integration possible! For example:
  - *an abstract model for the relational graphs: RDF*
  - *means to extract RDF information from XML (eg, XHTML) pages: GRDDL*
  - *means to add structured information to XHTML pages: RDFa*
  - *a query language adapted for the relational graphs: SPARQL*
  - *various technologies to characterize the relationships, categorize resources: RDFS (RDF Schemas), OWL (Web Ontology Language), SKOS, Rule Interchange Format*
    - depending on the complexity required, applications may choose among the different technologies
    - some of them may be relatively simple with simple tools (RDFS), whereas some require sophisticated systems (OWL, Rules)
  - *reuse of existing "ontologies" that others have produced (FOAF in our case)*
- Some of these technologies are stable, others are being developed

So where is the Semantic Web? (cont)



A real life data integration: Antibodies Demo

- Scenario: find the known antibodies for a protein in a specific species
- Combine four different data sources



Semantic Web data begins to accumulate on the Web

- Large datasets are accumulating. E.g.:
  - *IngentaConnect bibliographic metadata storage: over 200 million statements*
  - *RDF version of Wikipedia:*
    more than 47 million triplets, based also on SKOS, soon with a SPARQL interface
  - *tracking the US Congress: data stored in RDF (around 25 million triplets) with a SPARQL interface*
  - *"Département/canton/commune" structure of France published by the French Statistical Institute*
- Some mesaures claim that there are over 10^7
  Semantic Web documents… (ready to be integrated…)

Semantic Web Applications

Semantic Web ≠ an academic research only!

- SW has indeed a strong foundation in research results
- But remember:
  - *(1) the Web was born at CERN…*
  - *(2) …was first picked up by high energy physicists…*
  - *(3) …then by academia at large…*
  - *(4) …then by small businesses and start-ups…*
  - *(5) "big business" came only later!*
- network effect kicked in early…
- Semantic Web is now at #4, and moving to #5!

May start with small communities

- The needs of a deployment application area:
  - *have serious problem or opportunity*
  - *have the intellectual interest to pick up new things*
  - *have motivation to fix the problem*
  - *its data connects to other application areas*
  - *have an influence as a showcase for others*
- The high energy physics community played this role for the Web in the 90's

Some RDF deployment areas

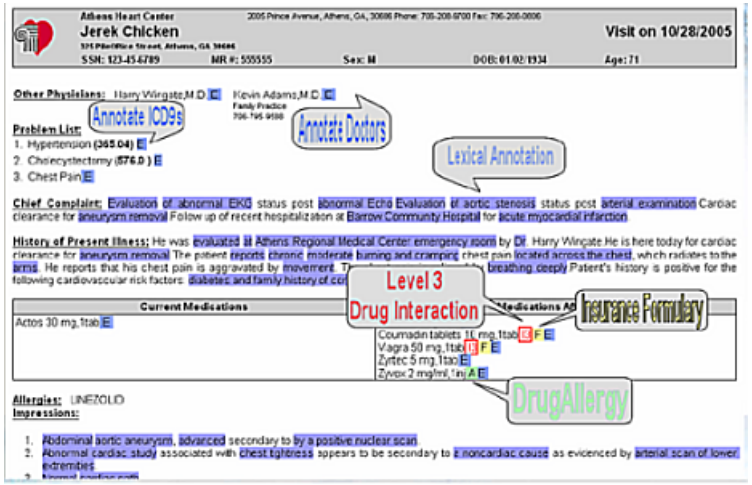
	Library metadata	Defense	Life sciences
Problem to solve?	single-domain integration	yes, serious data integration needs	yes, connections among genetics, proteomics, clinical trials, regulatory, …
Willingness to adopt?	yes: OCLC push and Dublin Core initiative	yes: funded early DAML (OWL) work	yes: intellectual level high, much modeling done already.
Motivation	light	strong	very strong
Links to	other library data	phone calls records, etc	chemistry, regulatory, medical, etc
Showcase?	very specialized	not at all	yes, model for other industries.

Some RDF deployment areas (cont)

- These are just examples
- Others are coming to the fore: *eGovernment*, energy sector (oil industry), financial services, …

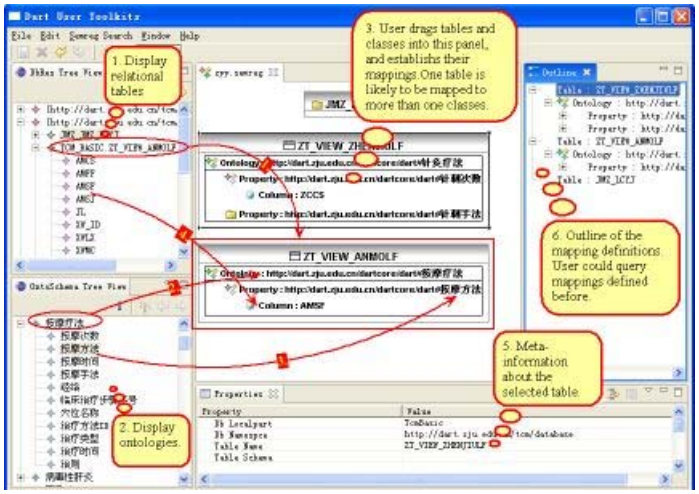
Applications are not always very complex…

- Eg: simple semantic annotations of patients' data greatly enhances communications among doctors
- What is needed: some simple ontologies, an RDFa/microformat type editing environment
- Simple but powerful!



Data Integration R&D

- Boeing, MITRE Corp., Elsevier, EU Projects like Sculpteur and Artiste, national projects like MuseoSuomi, DartGrid from Zhe Jiang University, …



Portals

- Vodafone's Live Mobile Portal
  - *search application (e.g. ringtone, game, picture) using RDF*
    - page views per download decreased 50%
    - ringtone up 20% in 2 months
- A number of other portal examples: Sun's White Paper Collections and System Handbook collections; Nokia's S60 support portal; Harper's Online magazine linking items via an internal ontology; Oracle's virtual press room; Opera's community site, Yahoo! Food,…
- Another example: semantic "harvester" of environmental agencies and information

Creative Commons

- To express rights of digital content on the Web
  - *legal constraints referred to in RDF, added to pages*
- There are specialized browsers, browser plugins
- More than 1,000,000 users worldwide (!)
  - *without knowing that they use RDF…*

Other Application Areas Come to the Fore

- Knowledge management
- Business intelligence
- Linking virtual communities
- Management of multimedia data (e.g., video and image depositories)
- Content adaptation and labeling (e.g., for mobile usage)
- etc

Conclusions

- The Semantic Web is there to integrate data on the Web
- The goal is the creation of a *Web of Data*

Thank you for your attention!

These slides are available on the Web: http://www.w3.org/2007/Talks/0202-Gijon-IH/

(slides are available in XHTML and PDF)