

Tutorial on the Semantic Web

Ivan Herman, W3C

(Last update: 5 November 2007)

> Introduction



> Towards a Semantic Web



- The current Web represents information using
 - natural language (English, Hungarian, Chinese,...)
 - graphics, multimedia, page layout
- Humans can process this easily
 - can deduce facts from partial information
 - can create mental associations
 - are used to various sensory information
 - (well, sort of... people with disabilities may have serious problems on the Web with rich media!)

> Towards a Semantic Web



- Tasks often require to combine data on the Web:
 - hotel and travel information may come from different sites
 - searches in different digital libraries
 - etc.
- Again, humans combine these information easily
 - even if different terminologies are used!

> However...



- However: machines are ignorant!
 - partial information is unusable
 - difficult to make sense from, e.g., an image
 - drawing analogies automatically is difficult
 - difficult to combine information automatically
 - is `<foo:creator>` same as `<bar:author>`?
 - how to combine different XML hierarchies?
 - ...

> Example: automatic airline reservation



- Your automatic airline reservation
 - knows about your preferences
 - builds up knowledge base using your past
 - can combine the local knowledge with remote services:
 - airline preferences
 - dietary requirements
 - calendaring
 - etc
- It communicates with remote information (i.e., on the Web!)
 - (M. Dertouzos: The Unfinished Revolution)

> Example: data(base) integration



- Databases are very different in structure, in content
- Lots of applications require managing several databases
 - 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)

> And the problem is real...



The screenshot displays three overlapping web browser windows illustrating database integration in neuroscience:

- CoCoDat (top):** Titled "Collection of Cortical Data - Mozilla Firefox", it shows a navigation menu with links like "CoCoDat", "DATABASES", "ORT", "EXAMPLES", "DOCUMENTS", "REFERENCES", and "CONTACTS". The main content area is titled "CoCoDat: Collation of Cortical [single neuron + neuronal microcircuitry] Data".
- NeuronDB (middle):** Titled "NeuronDB - Retinal photoreceptor - Overview (4) - Mozilla Firefox", it shows a sidebar with "Mode: Overview", "Region: Ded Dep Soma AH A T All Compartments", "Properties: Receptors Channels Transmitters All Properties", and "Interoperation: Gene and Chromosome Experimental Data". The main content area shows a diagram of a retinal photoreceptor with labels "A", "OS", and "IS".
- Cell Centered Database (bottom):** Titled "Cell Centered Database - Mozilla Firefox", it shows a navigation menu with links like "About", "Data", "Updates", "Tools", "Links", "Help", and "Search CCDB". The main content area is titled "CELL CENTERED DATABASE" and displays a table of data.

ID	Cell type	Structure	MPT	Raw image	Thumbnails
1	Medium Spiny Neuron	Dendritic Tree	Optical section series and mosaic		Reconstruction Segment
2	Purkinje Neuron	Dendritic Tree	optical section series		
3	Purkinje Neuron	Dendritic	optical section series		

> Example: digital libraries



- It means catalogs on the Web
 - librarians have known how to do that for centuries
 - goal is to have this on the Web, World-wide
 - extend it to multimedia data, too
- But it is more: software agents should also be librarians!
 - help you in finding the right publications

> Example: semantics of Web Services



- Web services technology is great
- But if services are ubiquitous, searching issue comes up, for example:
 - “find me the best differential equation solver”
 - “check if it can be combined with the XYZ plotter service”
- It is necessary to characterize the service
 - not only in terms of input and output parameters...
 - ...but also in terms of its semantics

> 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

> In what follows...



- We will use a simplistic example to introduce the main Semantic Web concepts
- We take, as an example area, data integration

> The 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 simplified bookstore data (dataset “A”)

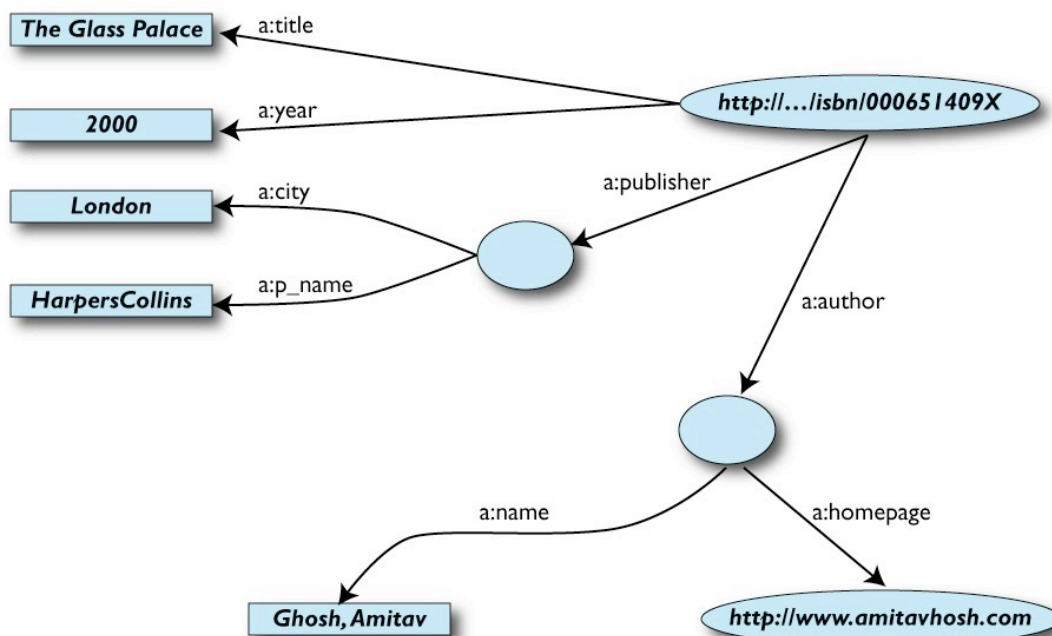


ID	Author	Title	Publisher	Year
ISBN 0-00-651409-X	id_xyz	The Glass Palace	id_qpr	2000

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

ID	Publ. Name	City
id_qpr	Harper Collins	London

> 1st: 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 HTML pages
 - extracting data from Excel sheets
 - etc.
- One can export *part* of the data

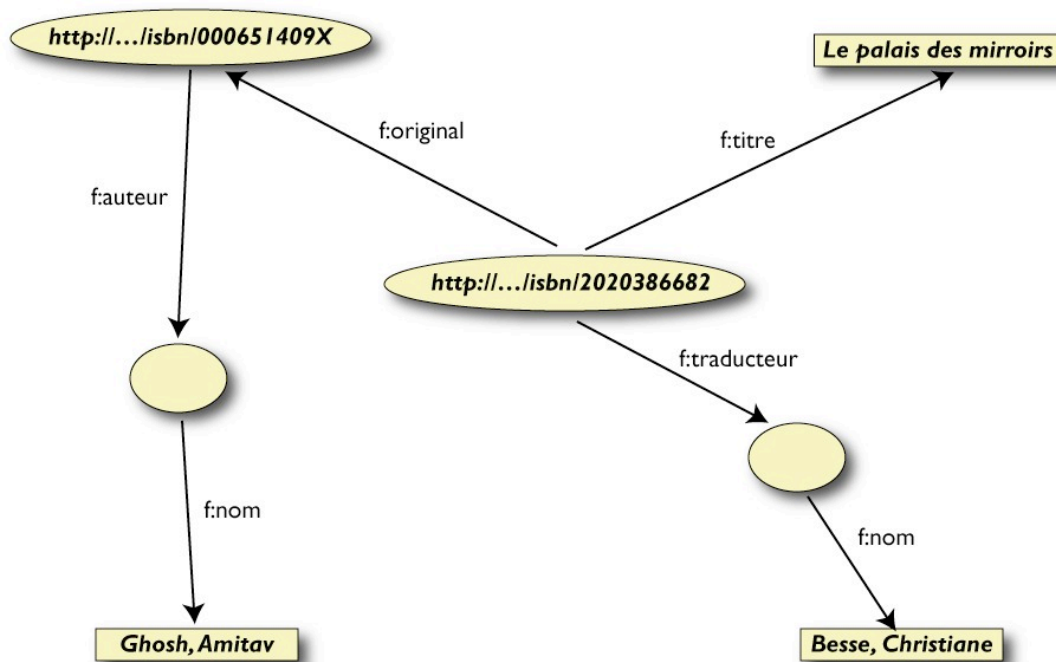
> Another bookshop data (dataset “F”)



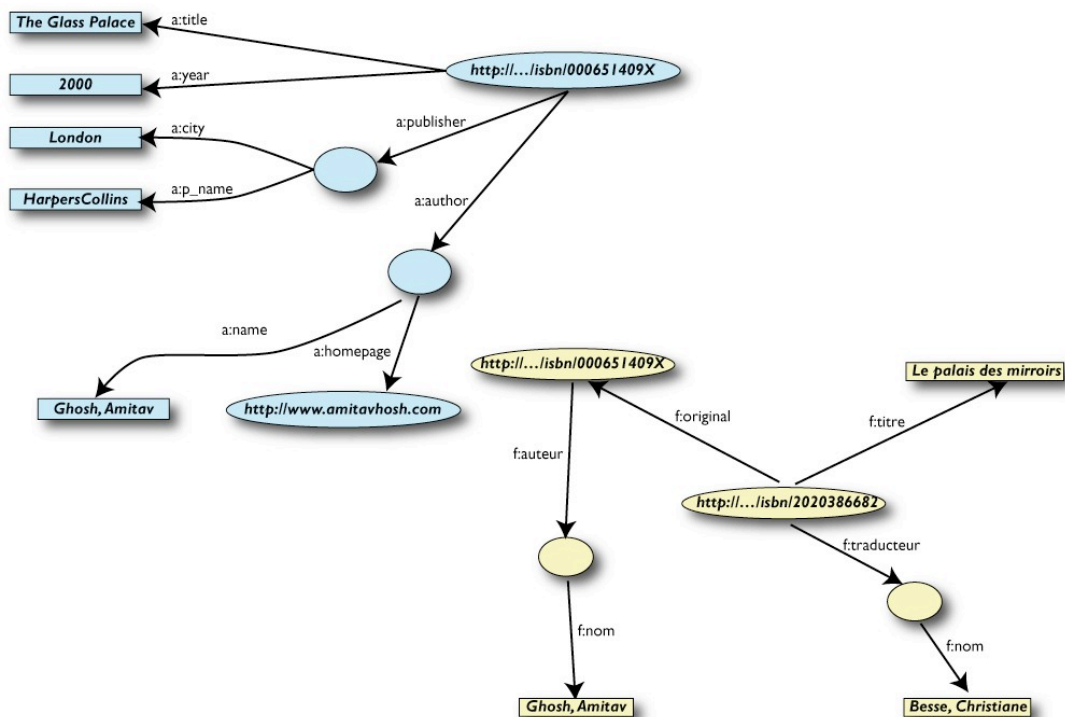
ID	Titre	Auteur	Traducteur	Original
ISBN 2020386682	Le Palais des miroirs	i_abc	i_qrs	ISBN 0-00-651409-X

ID	Nom
i_abc	Ghosh, Amitav
i_qrs	Besse, Christiane

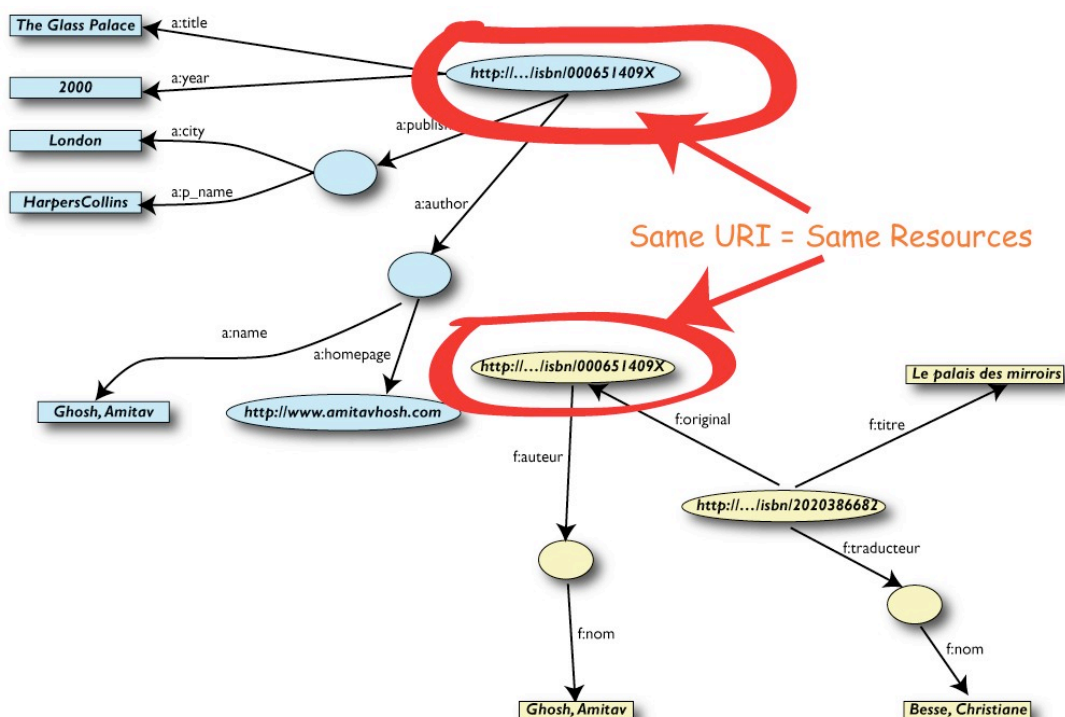
> 2nd: export your second set of data



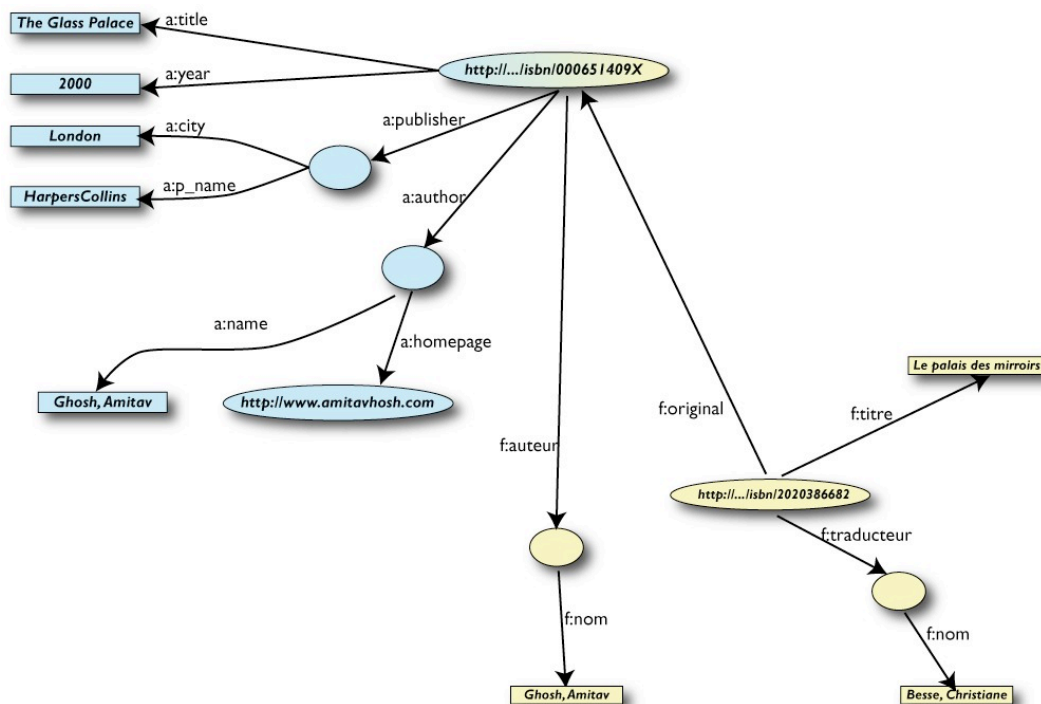
> 3rd: start merging your data



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



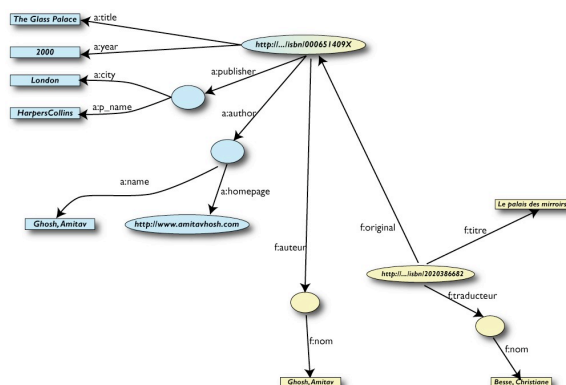
> 3rd: merge identical resources



> Start making queries...



- User of data “F” can now ask queries like:
 - « donnez-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 retrieved by merging with dataset “A”!

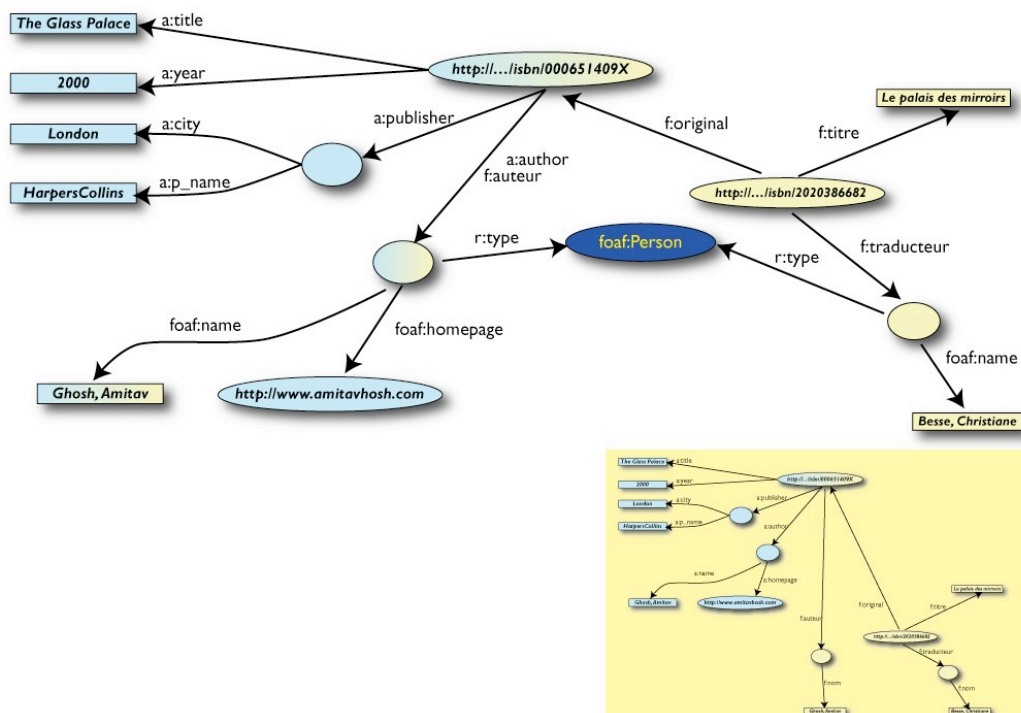


> However, more can be achieved...



- We “feel” that `a:author` and `f:auteur` should be the same
- But an automatic merge does 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

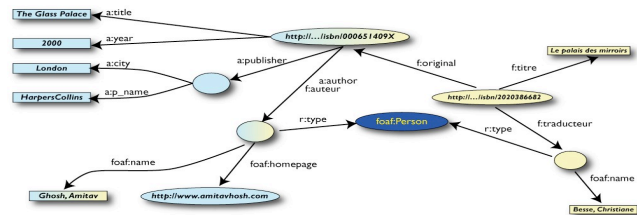
> 3rd revisited: use the extra knowledge



> Start making richer queries!



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

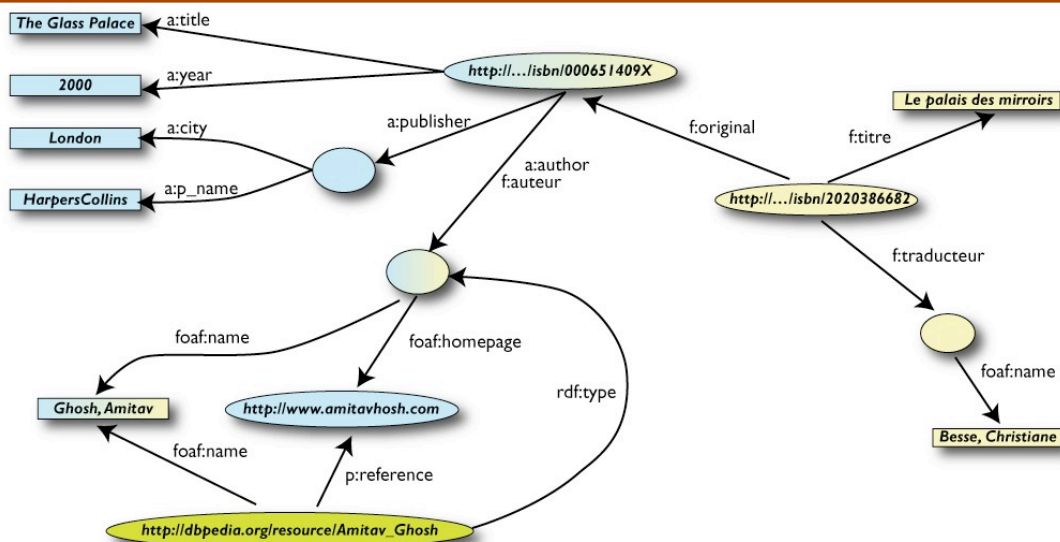


> Combine with different datasets

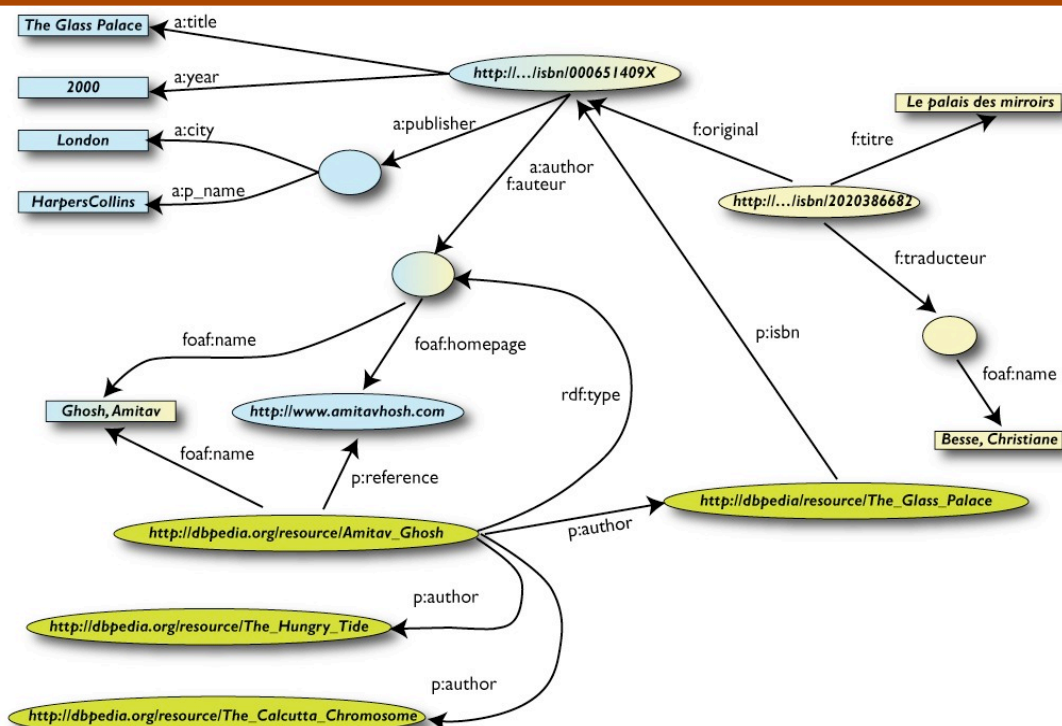


- Using, e.g., the “Person”, the dataset can be combined with other sources
- For example, data in Wikipedia can be extracted using dedicated tools
 - there is an active development to add some simple semantic “tag” to wikipedia entries (so called “Semantic Wiki”-s)
 - the “[dbpedia](#)” project can extract the “infobox” information from Wikipedia already...

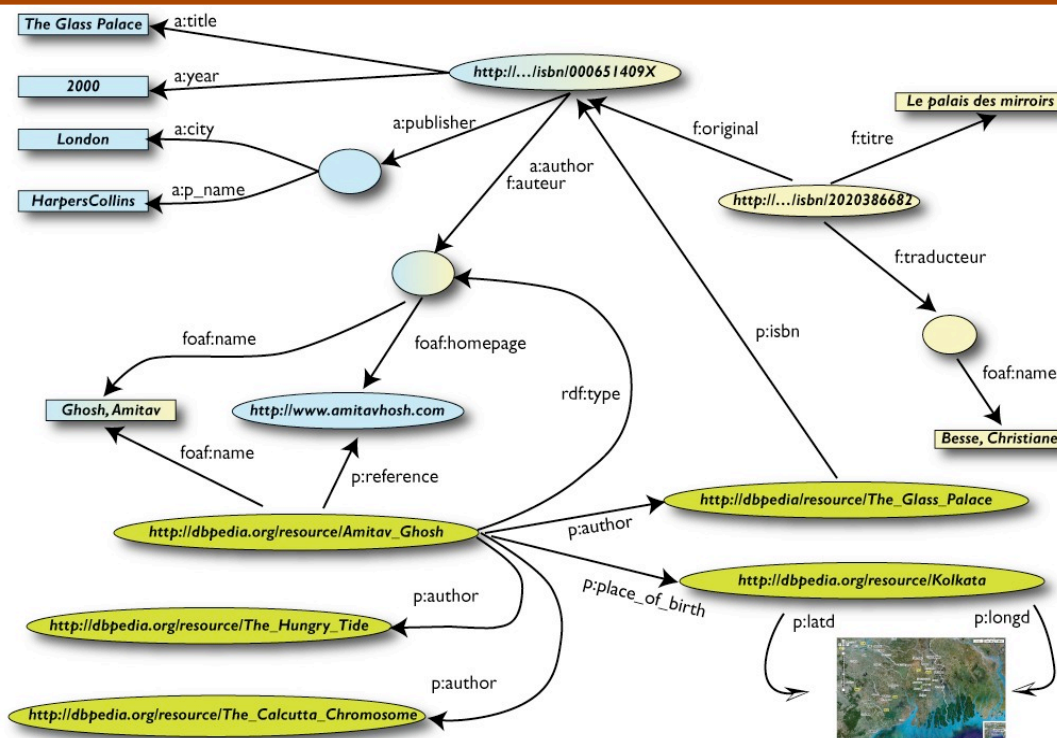
> Merge with Wikipedia data



> 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?



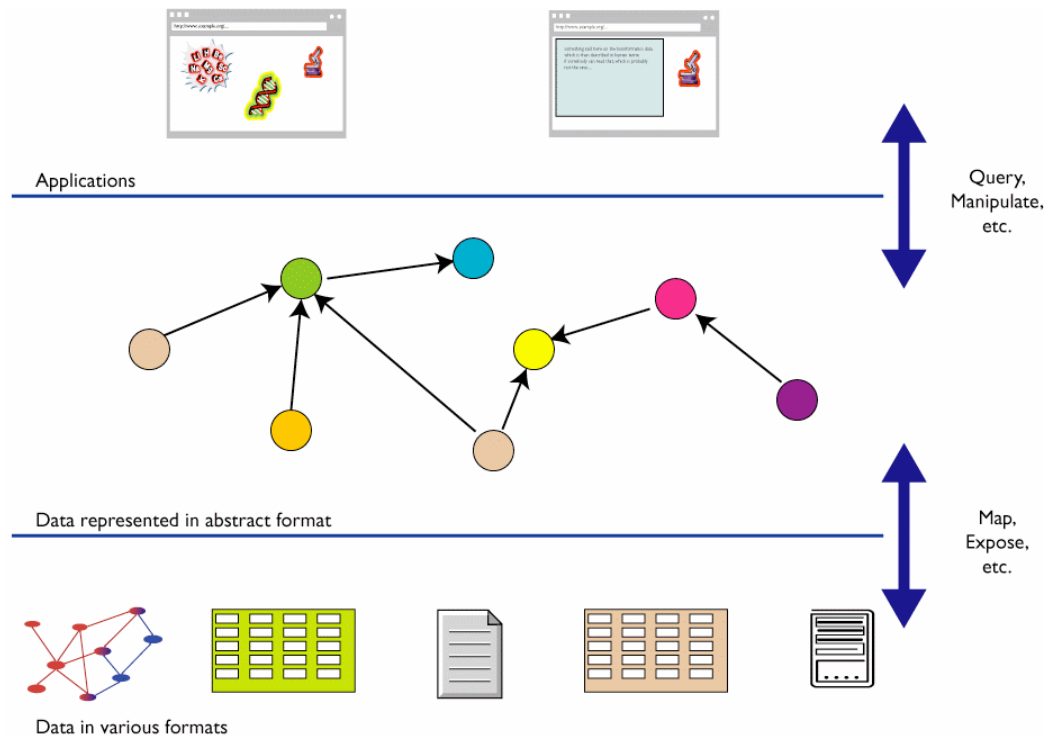
- We combined different datasets
 - all may be of different origin somewhere on the web
 - all may have different formats (mysql, excel sheet, XHTML, etc)
 - all may have different names for relations (e.g., multilingual)
- We could combine the data because some URI-s were identical (the ISBN-s in this 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
 - more geographical information
 - etc.
- This is where ontologies, thesauri, taxonomies, 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 schema's, 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 data sources