Introduction to Semantic Web

Ivan Herman, W3C

Deutsche Telekom Workshop Darmstadt, Germany 2009-12-14



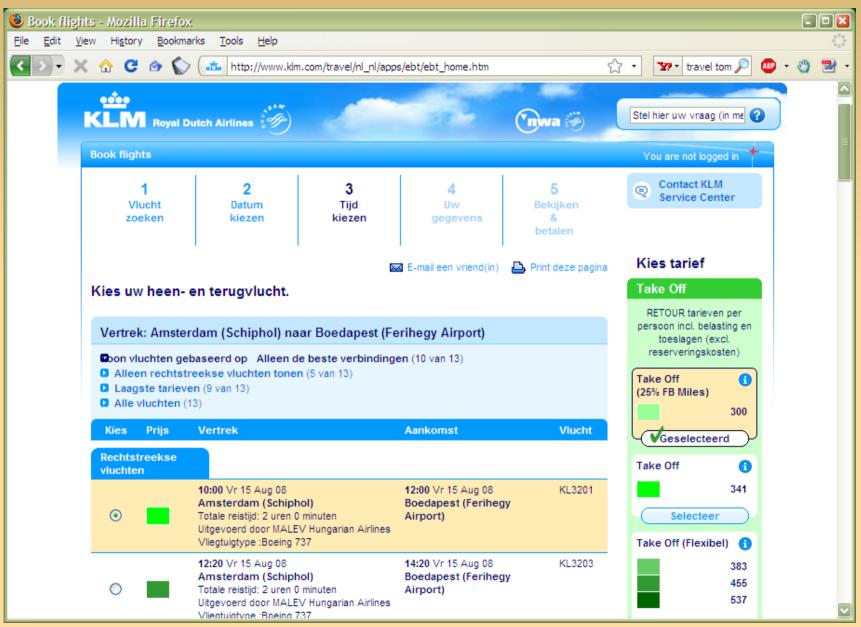
Let's organize a trip to Budapest using the Web!



You try to find a proper flight with ...

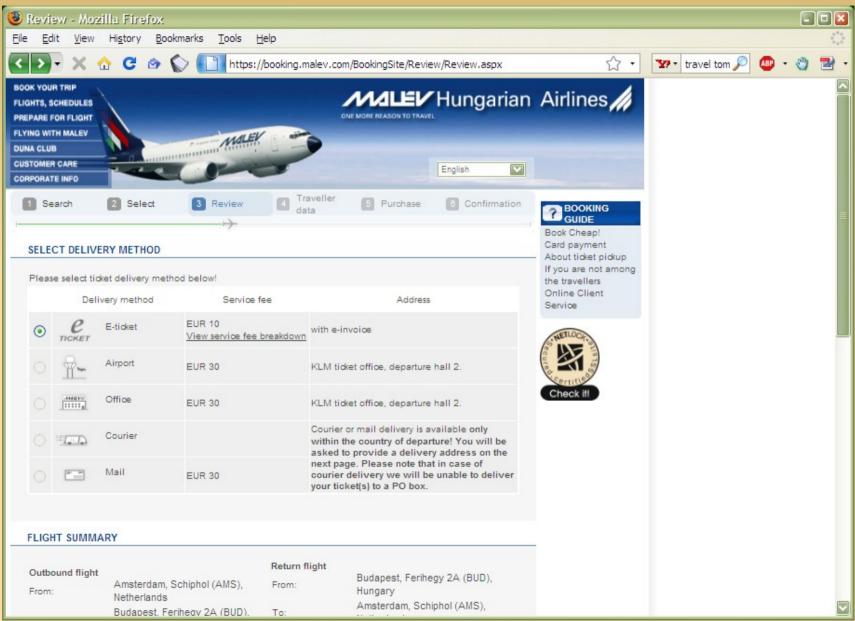


... a big, reputable airline, or ...



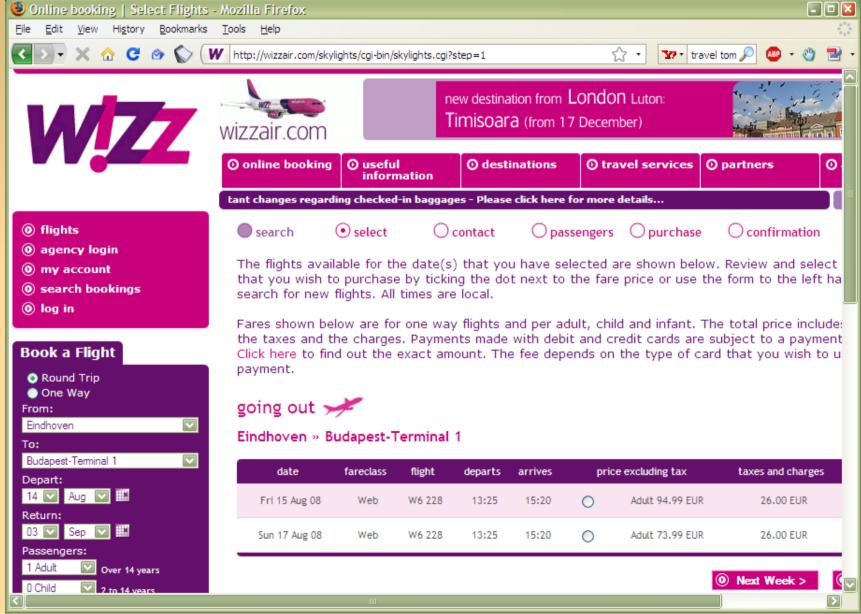


... the airline of the target country, or ...





... or a low cost one

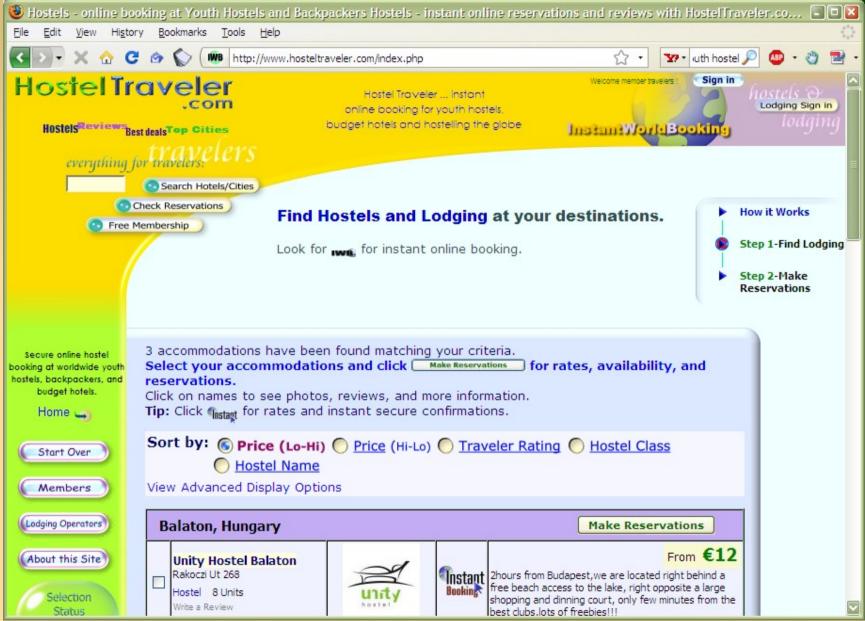




You have to find a hotel, so you look for...

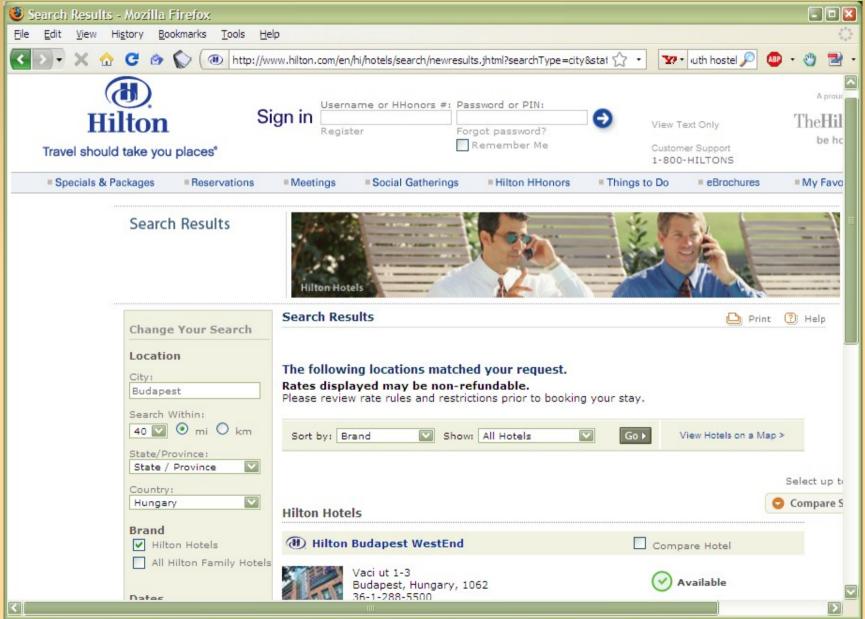


... a really cheap accommodation, or ...



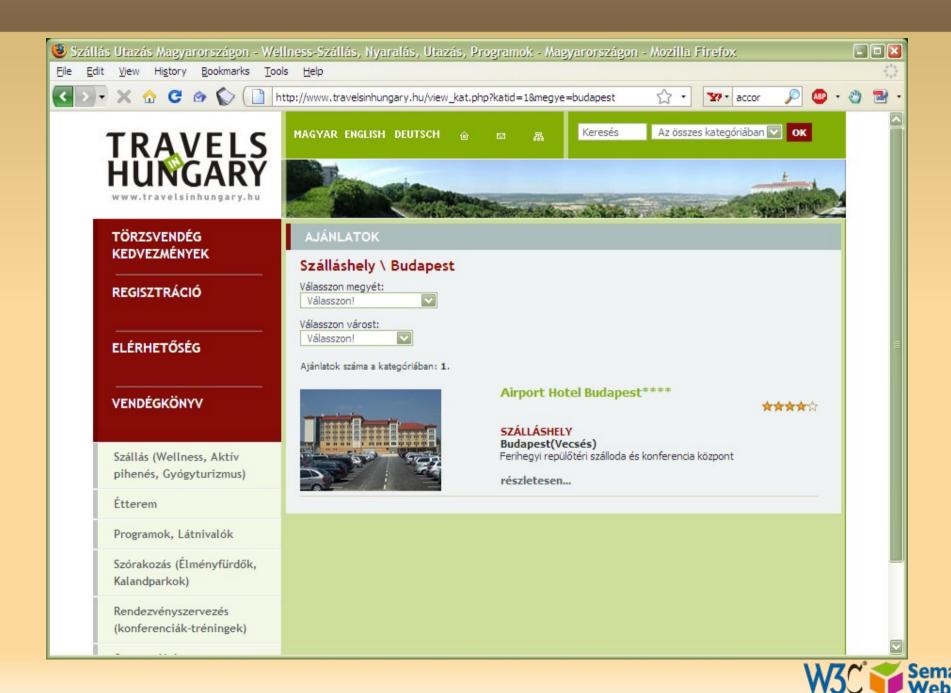


... or a really luxurious one, or ...





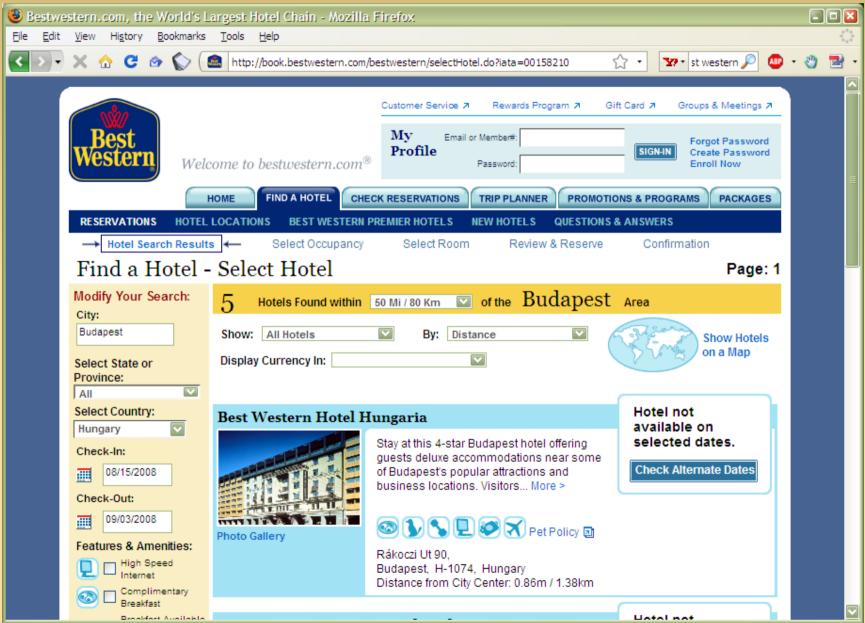
... an intermediate one ...



oops, that is no good, the page is in Hungarian that almost nobody understands, but...



... this one could work

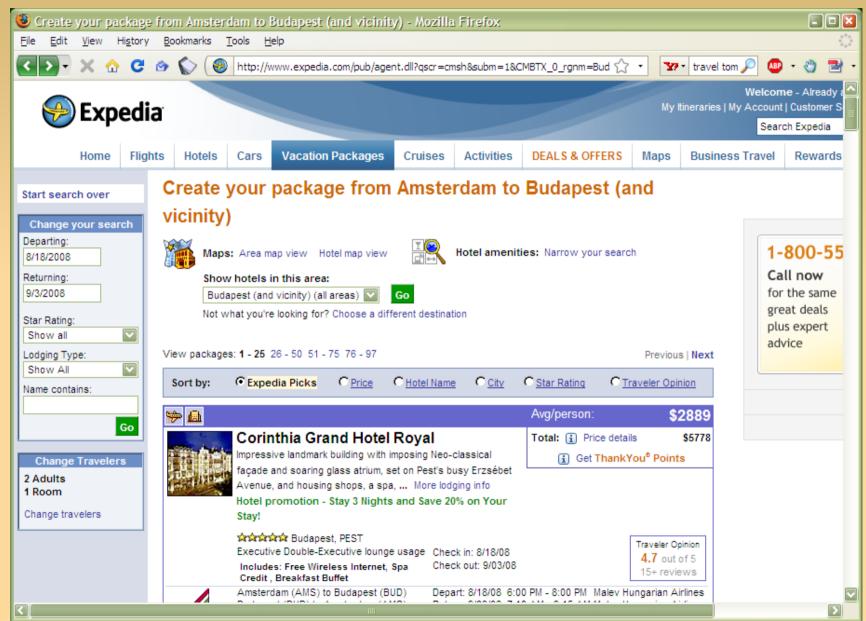




Of course, you could decide to trust a specialized site...

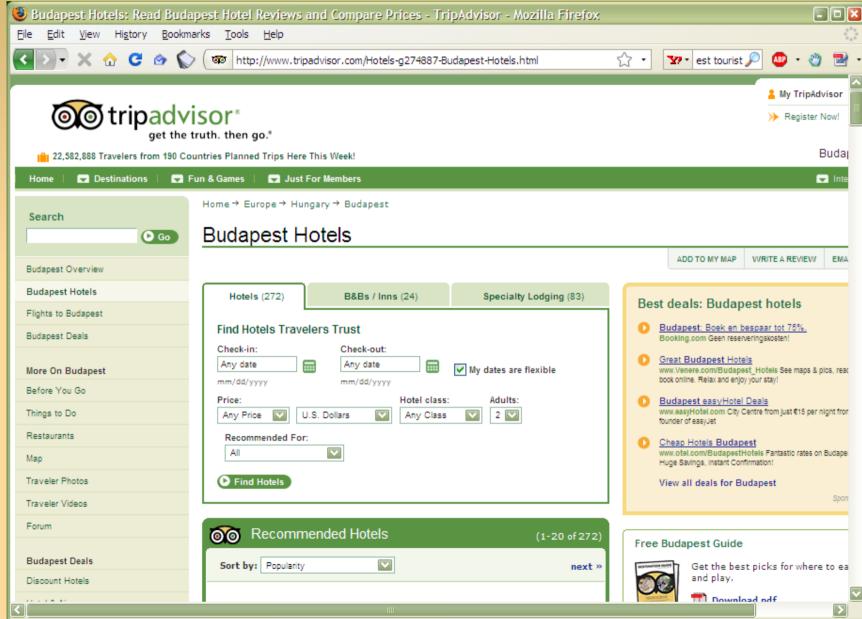


... like this one, or...





... or this one

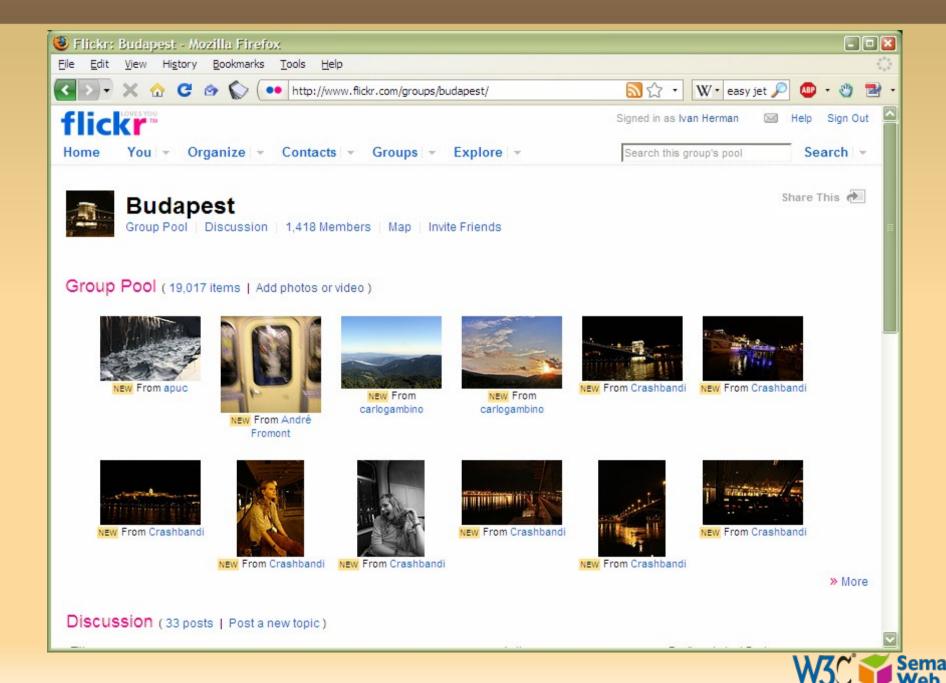




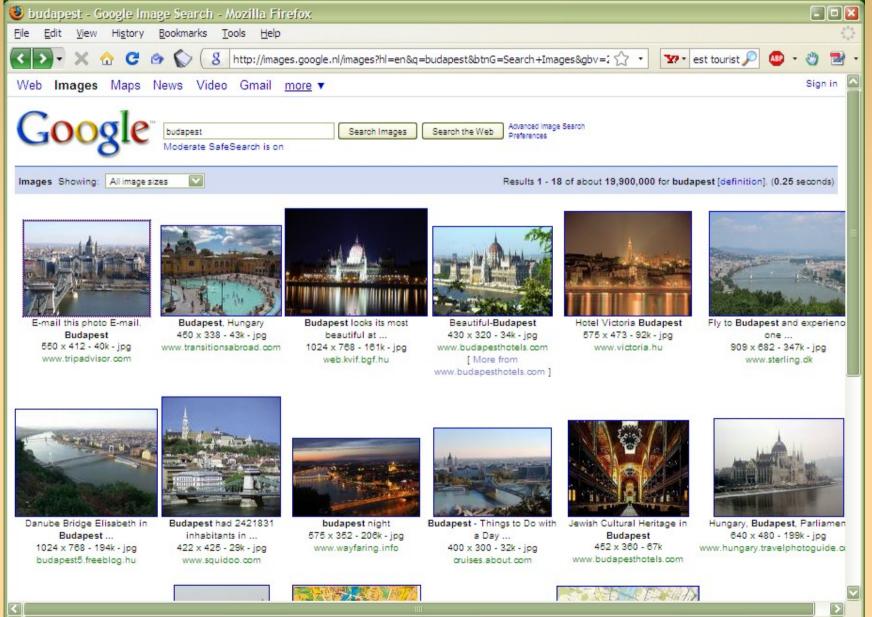
You may want to know something about Budapest; look for some photographs...



... on flickr ...

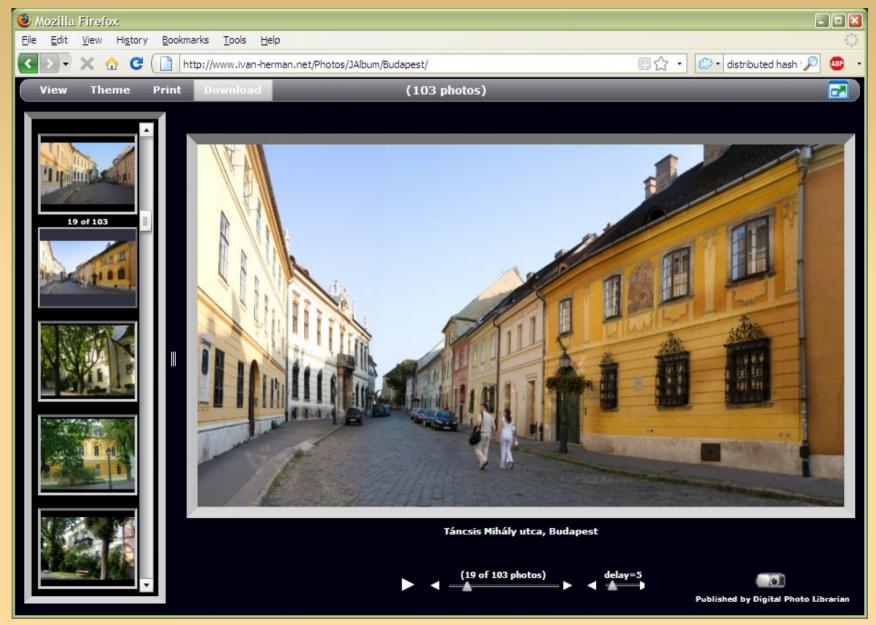


... on Google ...



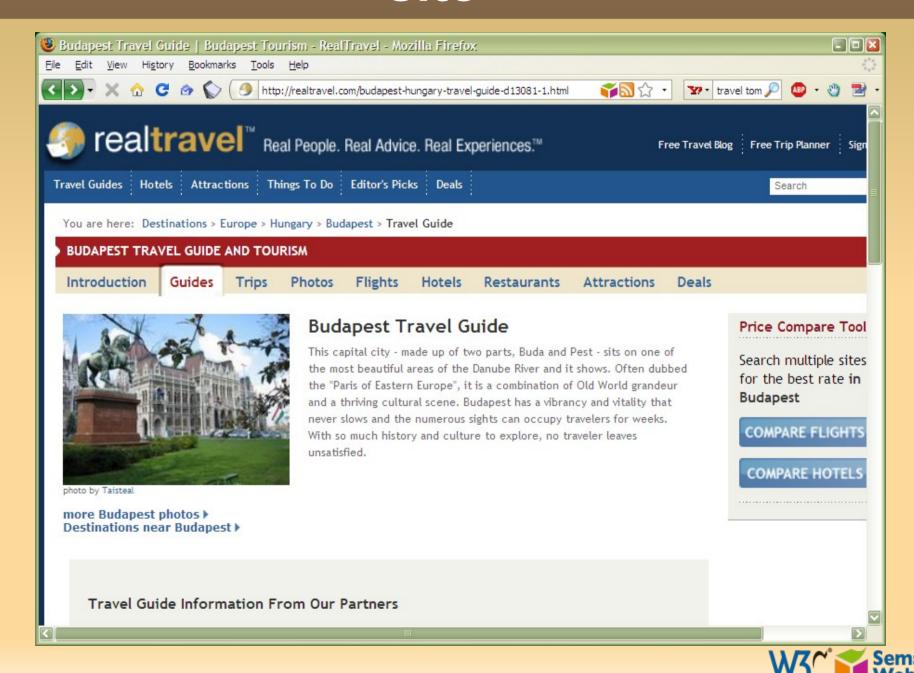


... or you can look at mine





but you can also look at a (social) travel²⁰ site



What happened here?

- You had to consult a large number of sites, all different in style, purpose, possibly language...
- You had to mentally integrate all those information to achieve your goals
- We all know that, sometimes, this is a long and tedious process!



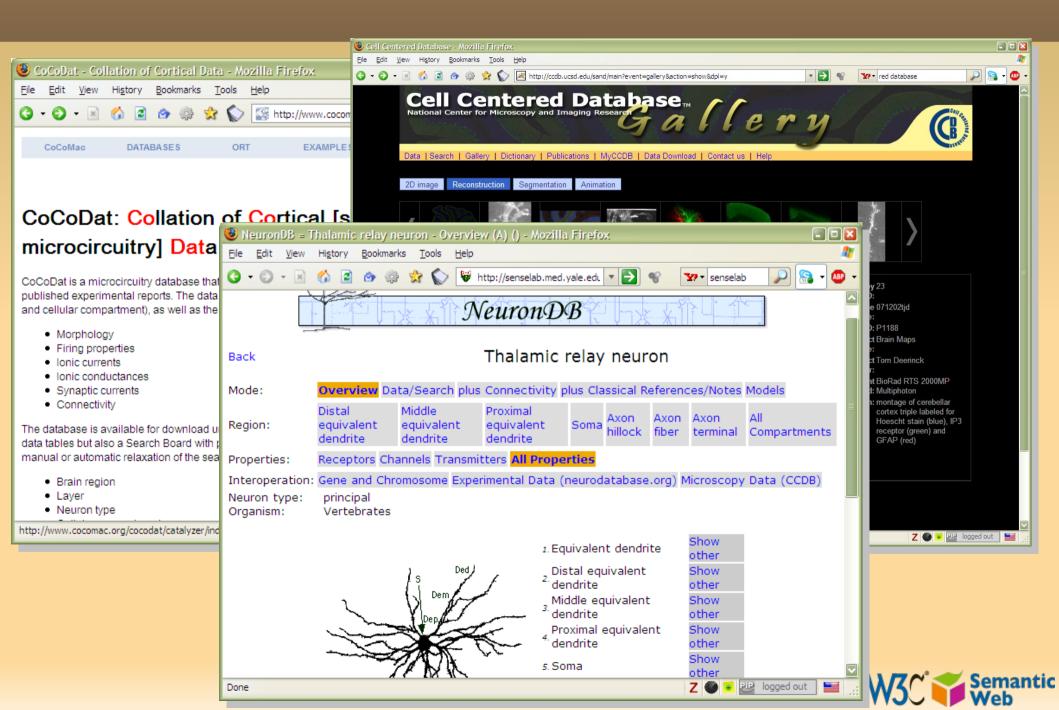
- All those pages are only tips of respective icebergs:
 - the real data is hidden somewhere in databases, XML files, Excel sheets, ...
 - you have only access to what the Web page designers allow you to see



- Specialized sites (Expedia, TripAdvisor) do a bit more:
 - they gather and combine data from other sources (usually with the approval of the data owners)
 - but they still control how you see those sources
- But sometimes you want to personalize: access the original data and combine it yourself!



Here is another example...



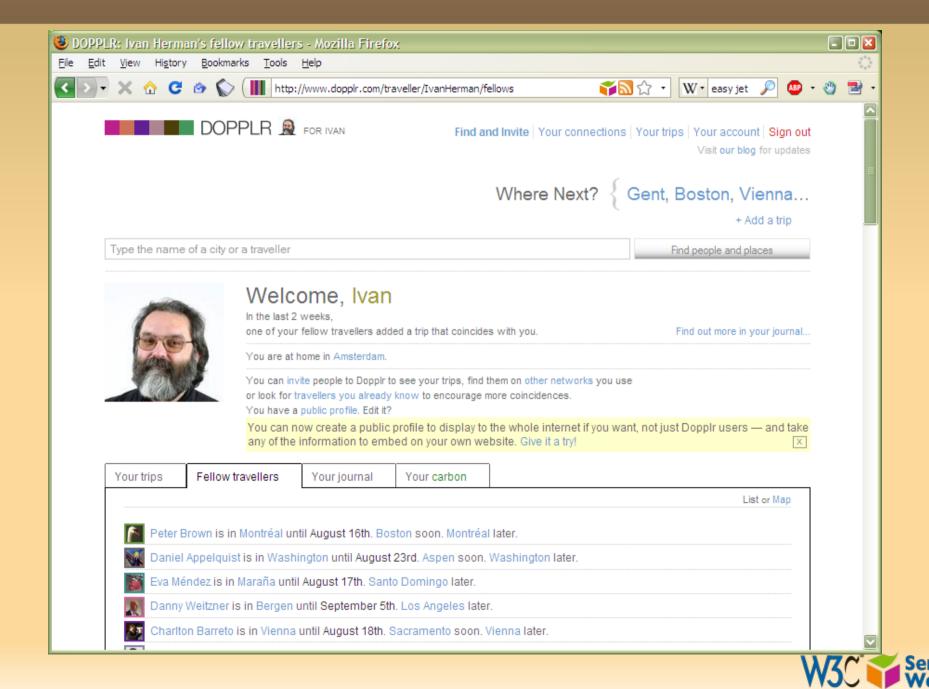
 Companies may have to hire a person to answer questions based on those (public!) databases!



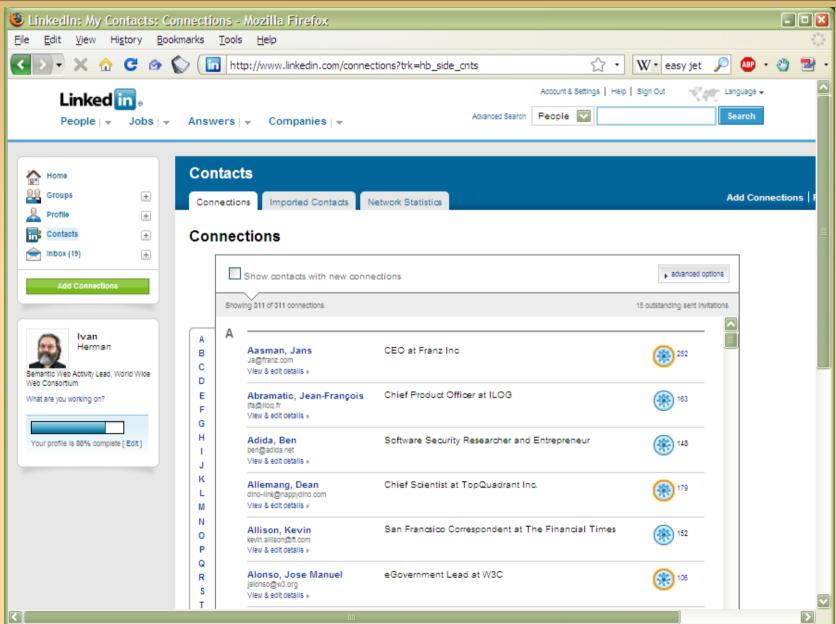
Another example: social sites. I have a list of "friends" by...



... Dopplr,

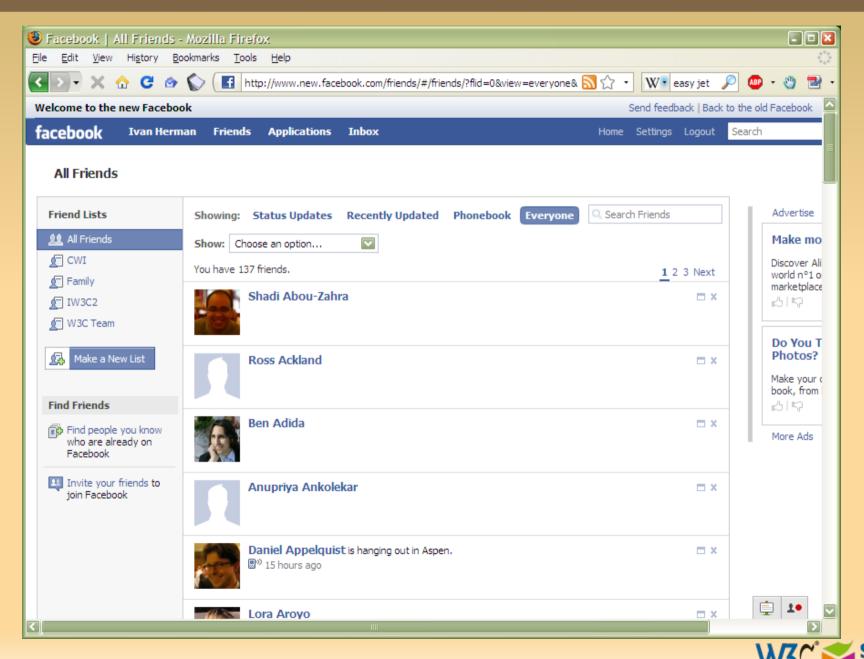


... LinkedIn,





... and, of course, the ubiquitous Facebook



- I had to type in and connect with friends again and again for each site independently
- This is even worse then before: I feed the icebergs, but I still do not have an easy access to data...



What would we like to have?

- Use the data on the Web the same way as we do with documents:
 - be able to link to data (independently of their presentation)
 - use that data the way I want (present it, mine it, etc)
 - agents, programs, scripts, etc, should be able to interpret part of that data



Put it another way....

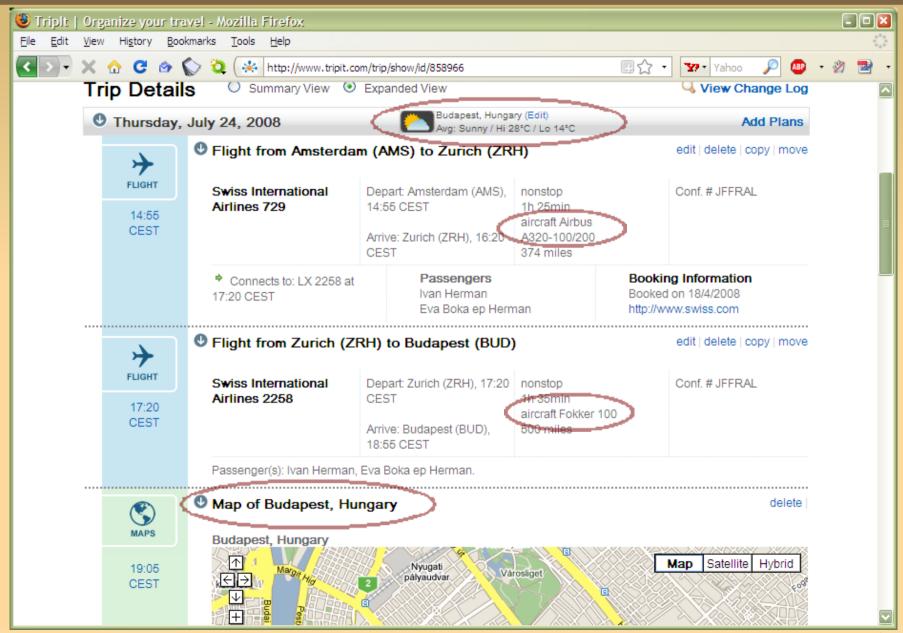
- We would like to extend the current Web to a "Web of data":
 - allow for applications to exploit the data directly



But wait! Isn't what mashup sites are already doing?



A "mashup" example:





- In some ways, yes, and that shows the huge power of what such Web of data provides
- But mashup sites are forced to do very ad-hoc jobs
 - various data sources expose their data via Web Services
 - each with a different API, a different logic, different structure
 - these sites are forced to reinvent the wheel many times because there is no standard way of doing things



Put it another way (again)...

 We would like to extend the current Web to a <u>standard</u> way for a "Web of data"



But what does this mean?

- What makes the current (document) Web work?
 - people create different documents
 - they give an address to it (ie, a URI) and make it accessible to others on the Web



An example: Steven's site on Amsterdam[®] (done for some visiting friends)

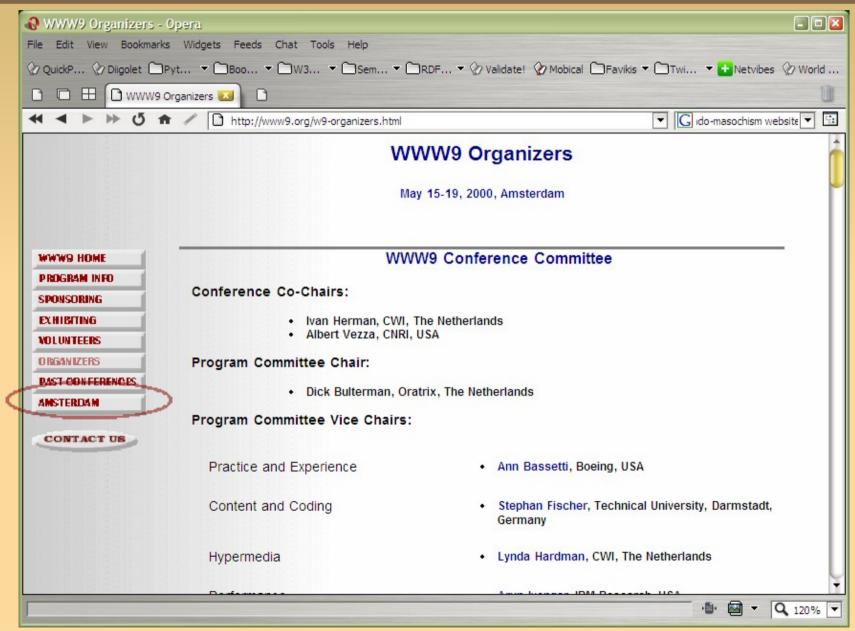


Then some magic happens...

- Others discover the site and they link to it
- The more they link to it, the more important and well known the page becomes
 - remember, this is what, eg, Google exploits!
- This is the "Network effect": some pages become important, and others begin to rely on it <u>even if the</u> <u>author did not expect it...</u>

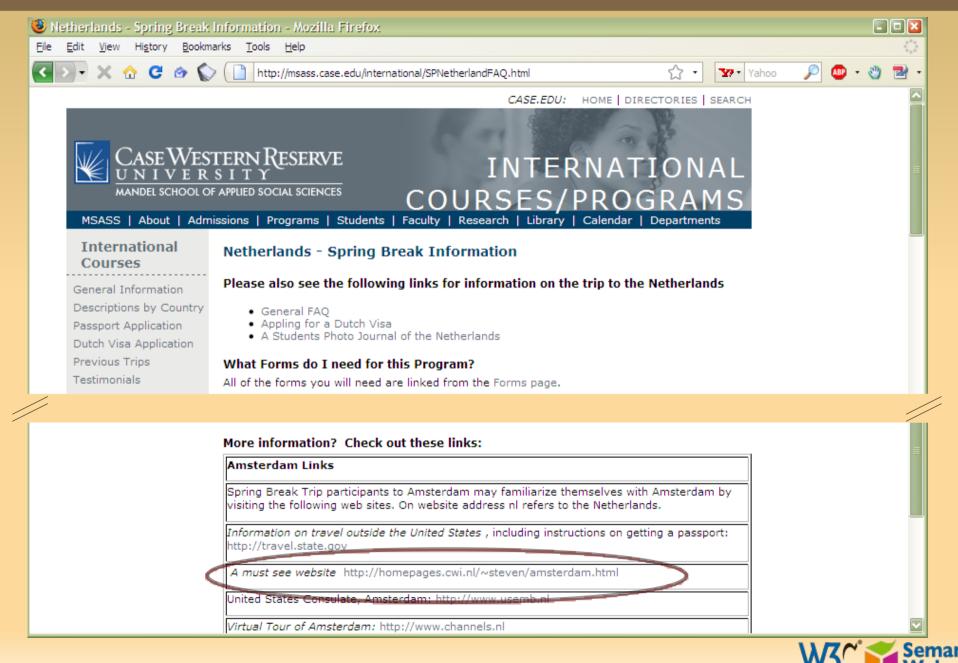


This could be expected...





but this one, from the other side of the Globe, was not...



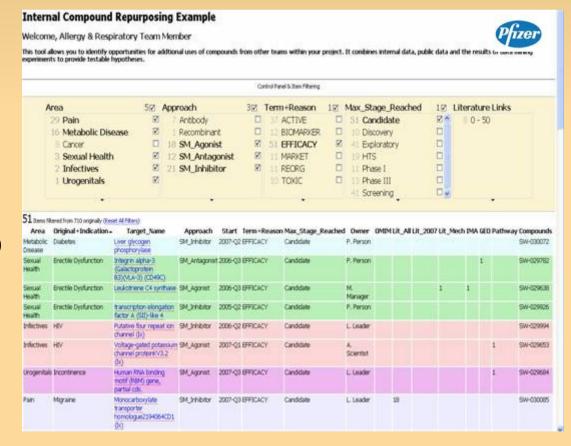
What would that mean for a Web of Data?

- Lessons learned: we should be able to:
 - "publish" the data to make it known on the Web
 - standard ways should be used instead of ad-hoc approaches
 - the analogous approach to documents: give URI-s to the data
 - make it possible to "link" to that URI from other sources of data (not only Web pages)
 - ie, applications should not be forced to make targeted developments to access the data
 - generic, standard approaches should suffice
 - and let the network effect work its way...



Example: combine data from experiments

- A drug company has huge amount of old experimental data on its Intranet
- Data in different formats (XML, databases, ...)
- To reuse them:
 - make the important facts available on the Web via standards
 - use off-the-shelf tool to integrate, display, search

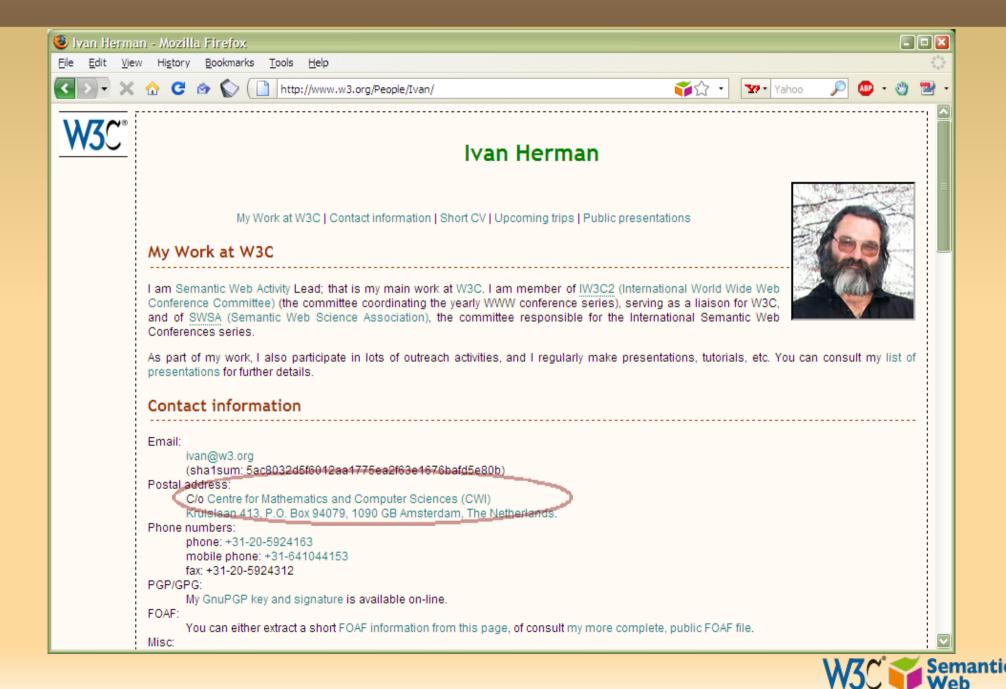


But it is a little bit more complicated

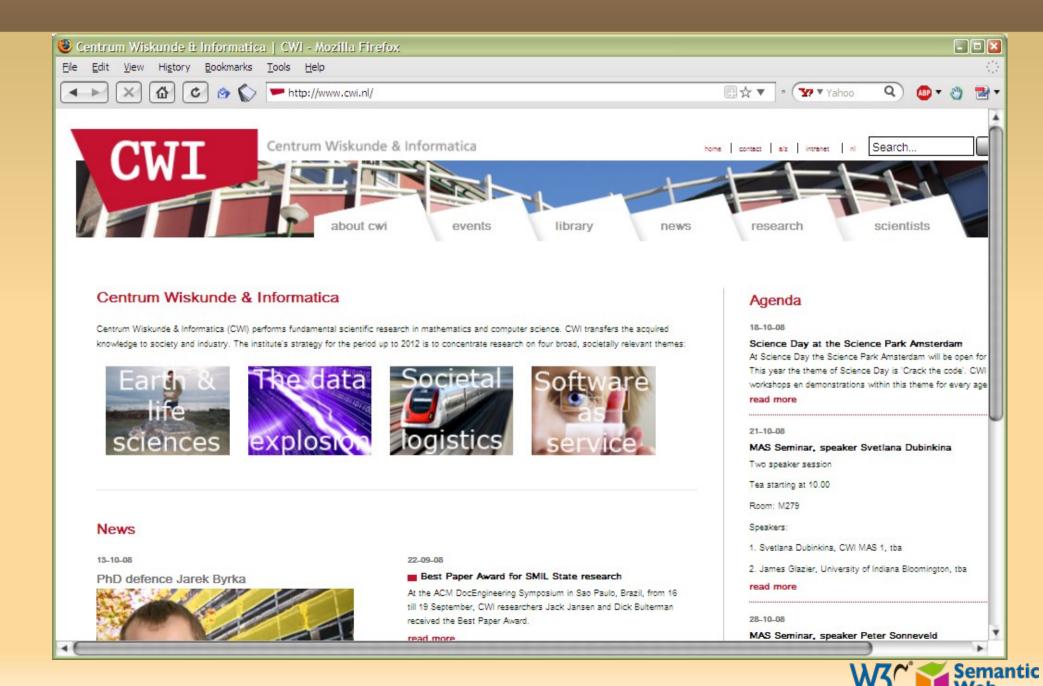
- On the traditional Web, humans are implicitly taken into account
- A Web link has a "context" that a person may use



Eg: address field on my page:



... leading to this page



- A human understands that this is my institution's home page
- He/she knows what it means (realizes that it is a research institute in Amsterdam)
- On a Web of Data, something is missing; machines can't make sense of the link alone



New lesson learned:

- extra information ("label") must be added to a link: "this links to my institution, which is a research institute"
- this information should be machine readable
- this is a characterization (or "classification") of both the link and its target
- in some cases, the classification should allow for some limited "reasoning"



Let us put it together

- What we need for a Web of Data:
 - use URI-s to publish data, not only full documents
 - allow the data to link to other data
 - characterize/classify the data and the links (the "terms") to convey some extra meaning
 - and use standards for all these!



So What is the Semantic Web?



It is a collection of standard technologies to realize a Web of Data



In what follows...

 We will use a simplistic example to introduce the main technical concepts of the Semantic Web



The rough structure of data integration

- 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



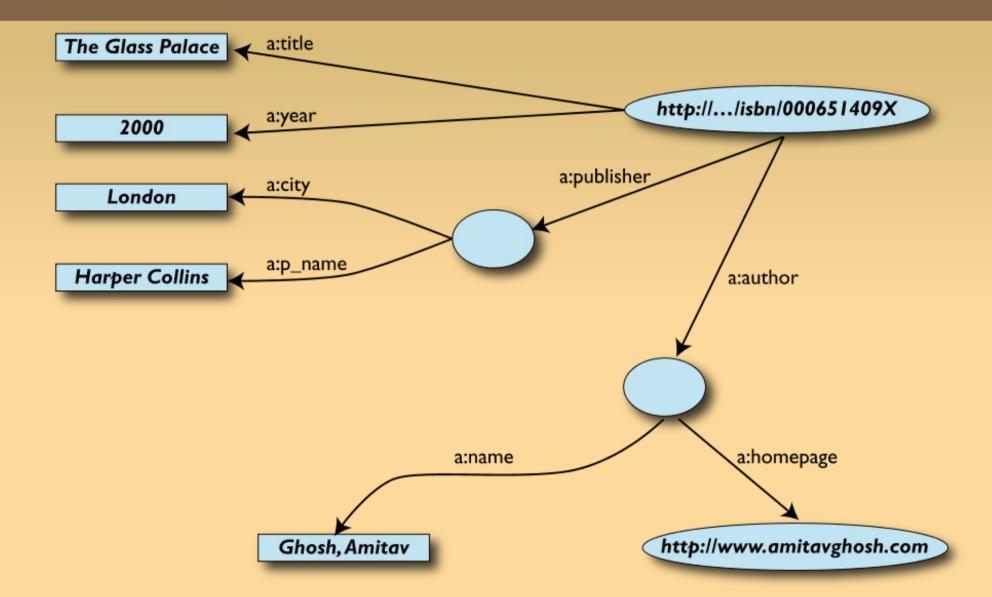
A <u>simplified</u> bookstore data (dataset "A")

ID	Author	Title	Publisher	Year
ISBN0-00-651409-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







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 <u>not</u> 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 <u>part</u> of the data

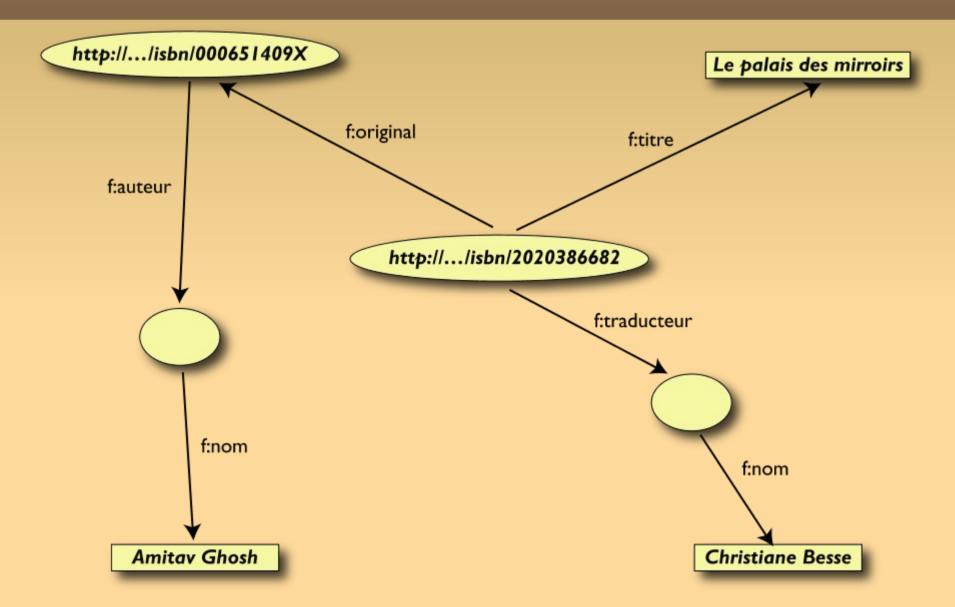


Another bookstore data (dataset "F")

	A	В	D	E
1	ID	Titre	Traducteur	Original
	ISBN0 2020386682	Le Palais des miroirs	A13	ISBN-0-00-651409-X
2		111110113		
3				
6	ID	Auteur		
7	ISBN-0-00-651409-X	A12		
11	Nom			
12	Ghosh, Amitav			
13	Besse, Christianne			

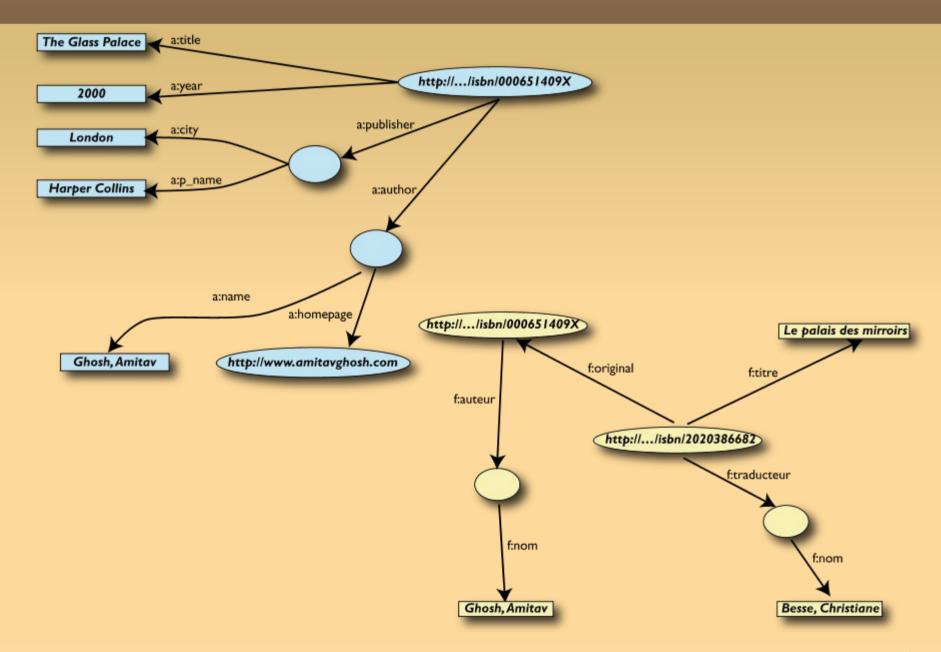


2nd: export your second set of data



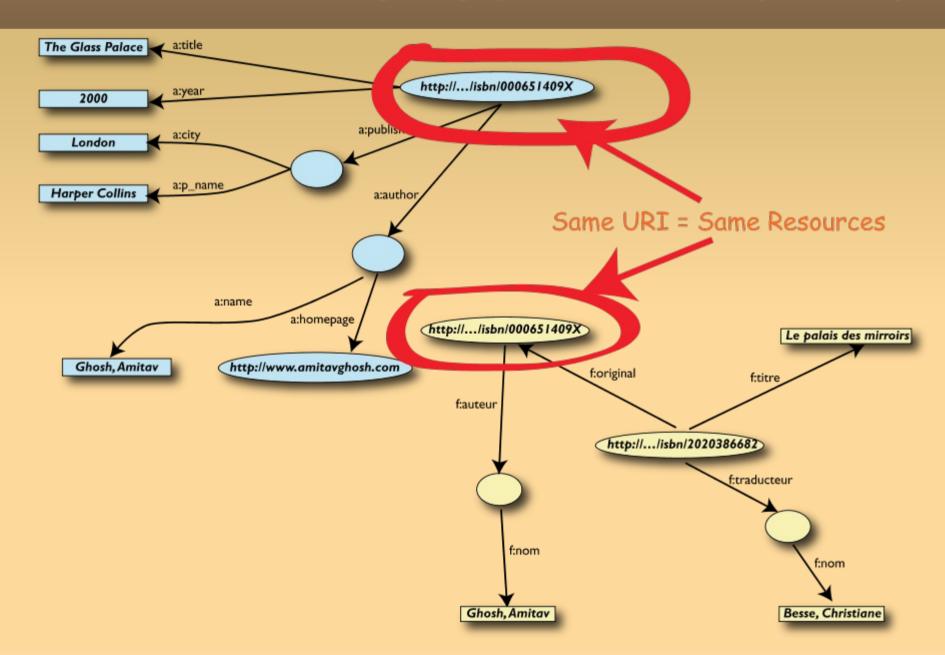


3rd: start merging your data



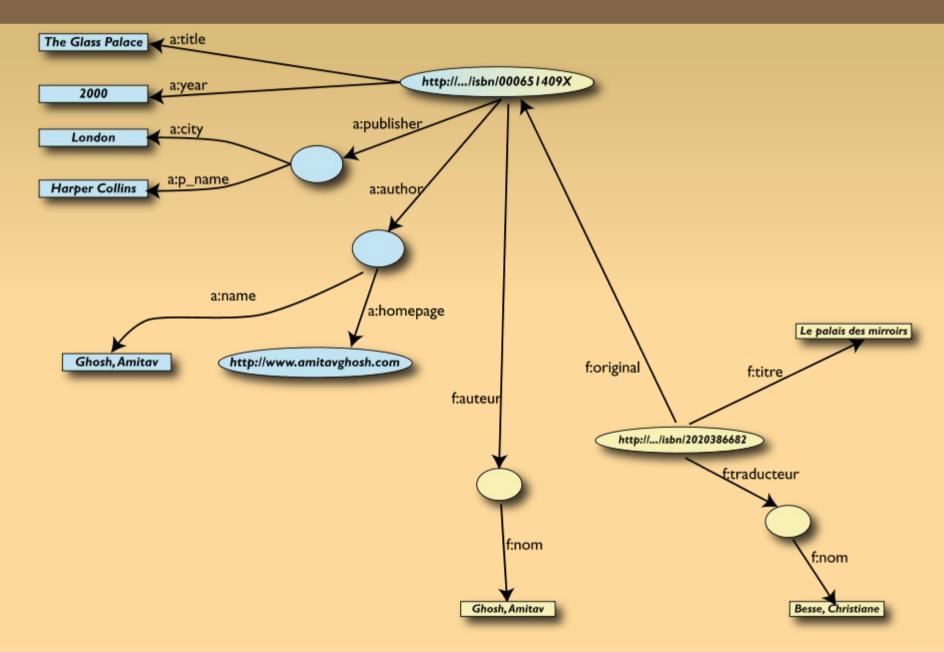


3rd: start merging your data (cont.)





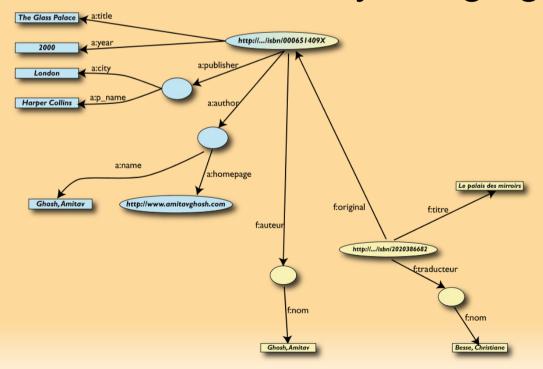
3rd: merge identical resources





Start making queries...

- User of data "F" can now ask queries like:
 - "give me the title of the original"
 - well, ... « donnes-moi le titre de l'original »
- This information is not in the dataset "F"....
- ...but can be retrieved by merging with dataset "A"!



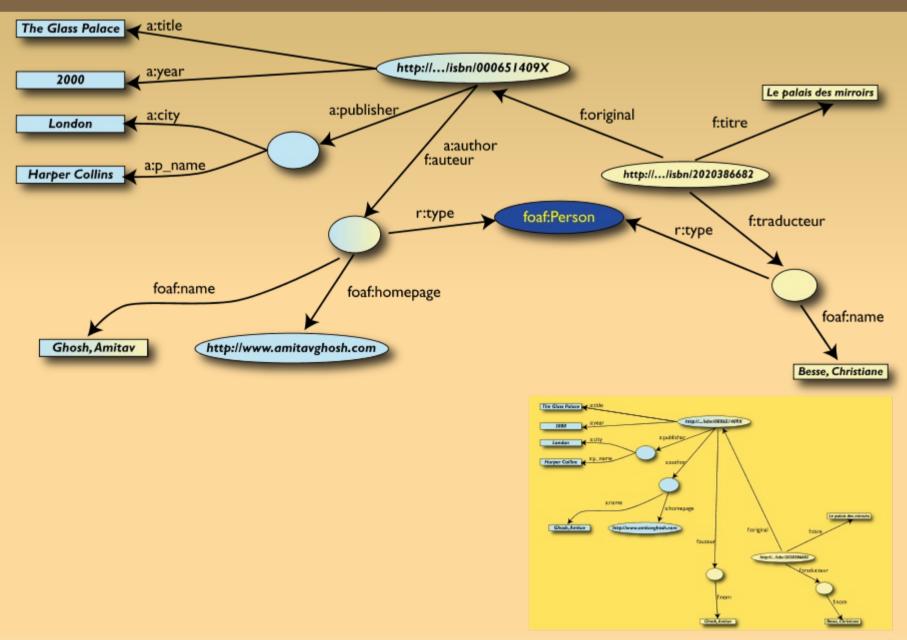


However, more can be achieved...

- We "feel" that a:author and f:auteur should be the same
- But an automatic merge 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



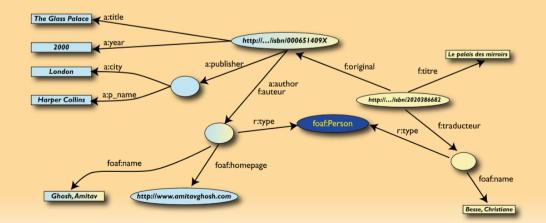
3rd revisited: use the extra knowledge





Start making richer queries!

- User of dataset "F" can now query:
 - "donnes-moi la page d'accueil de l'auteur de l'originale"
 - well... "give me the home page of the original's 'auteur'"
- 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"



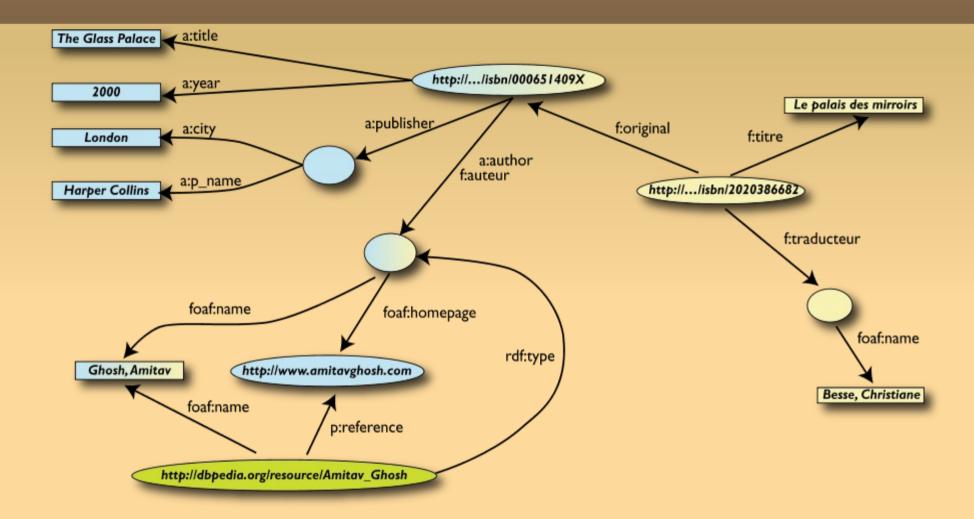


Combine with different datasets

- Using, e.g., the "Person", the dataset can be combined with other sources
- For example, data in Wikipedia can be extracted using dedicated tools
 - e.g., the "dbpedia" project can extract the "infobox" information from Wikipedia already…

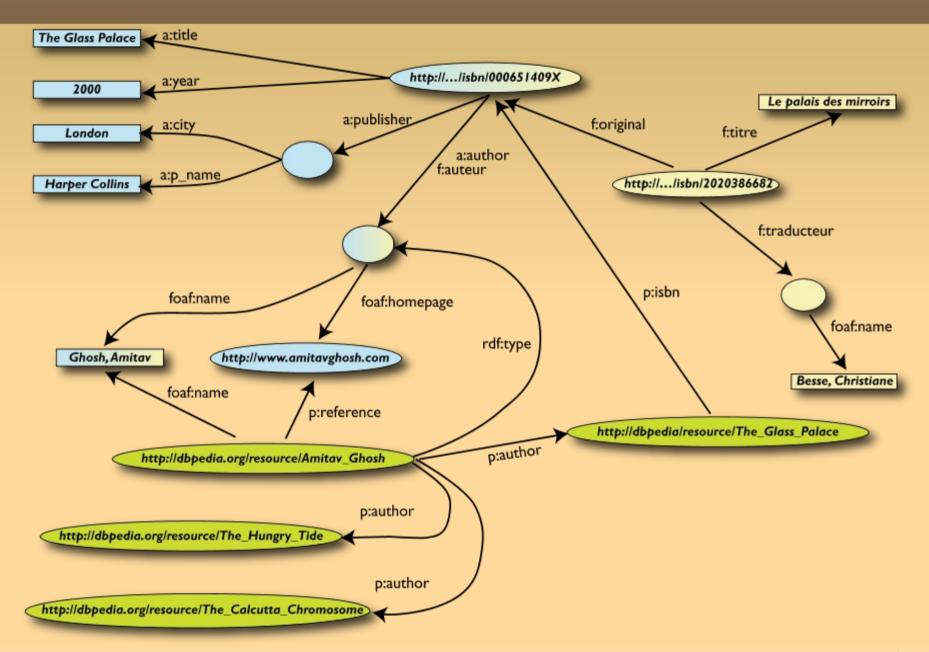


Merge with Wikipedia data



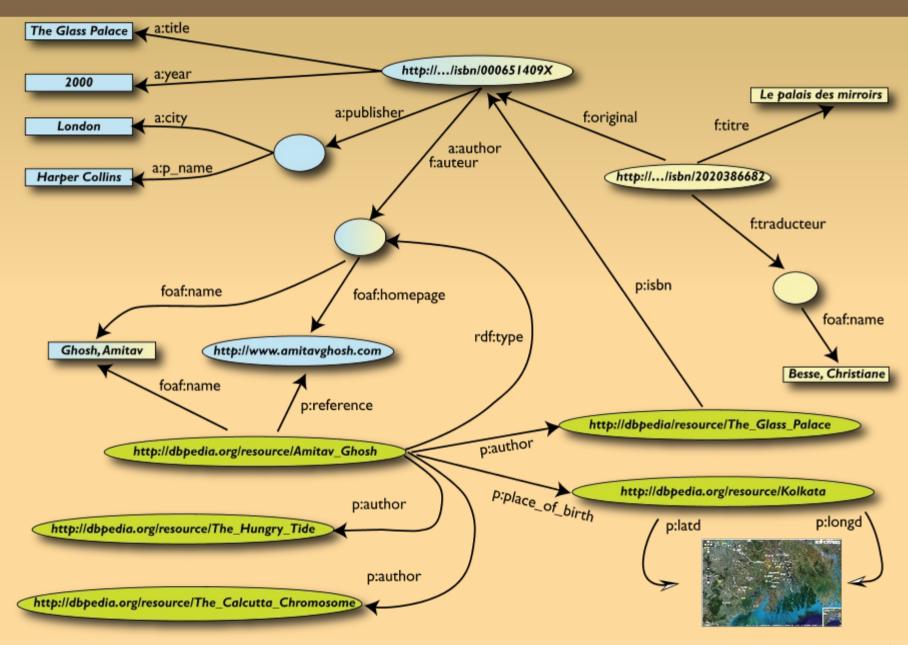


Merge with Wikipedia data





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 rigour (e.g., naming the relationships) is necessary so that machines could do this, too



What did we do?

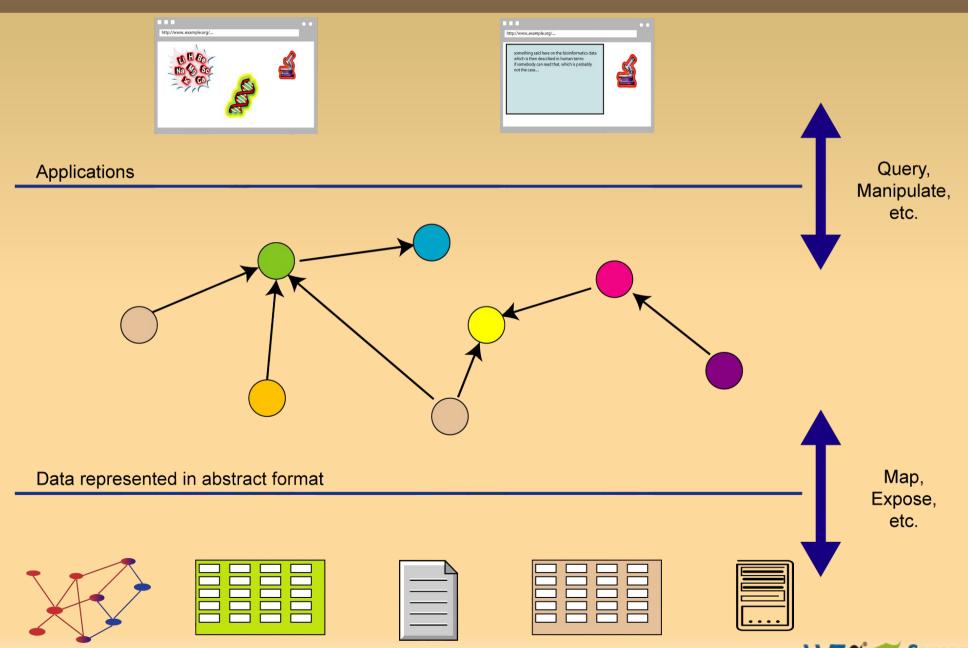
- We combined different datasets that
 - are somewhere on the web
 - are of different formats (mysql, excel sheet, XHTML, etc)
 - have different names for relations
- We could combine the data because some URI-s were identical (the ISBN-s in this case)
- We could add some simple additional information, 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 types of library data
 - geographical information
 - etc.
- This is where <u>ontologies</u>, extra <u>rules</u>, etc, come in
 - ontologies/rule sets can be relatively simple and small, or huge, or anything in between...
- Even more powerful queries can be asked as a result



What did we do? (cont)



Data in various formats

W3C Semantic

The technical building blocks...

- An abstract model for the relational graphs: RDF
- Convert/assign RDF graphs to existing data: GRDDL, RDFa, POWDER, R2RML
- A query language adapted for the relational graphs:
 SPARQL













The technical building blocks...

- Characterize the relationships, categorize resources: RDFS, OWL, SKOS, Rules
 - applications may choose among the different technologies
- Reuse of existing "ontologies" that others have produced (FOAF in our case)



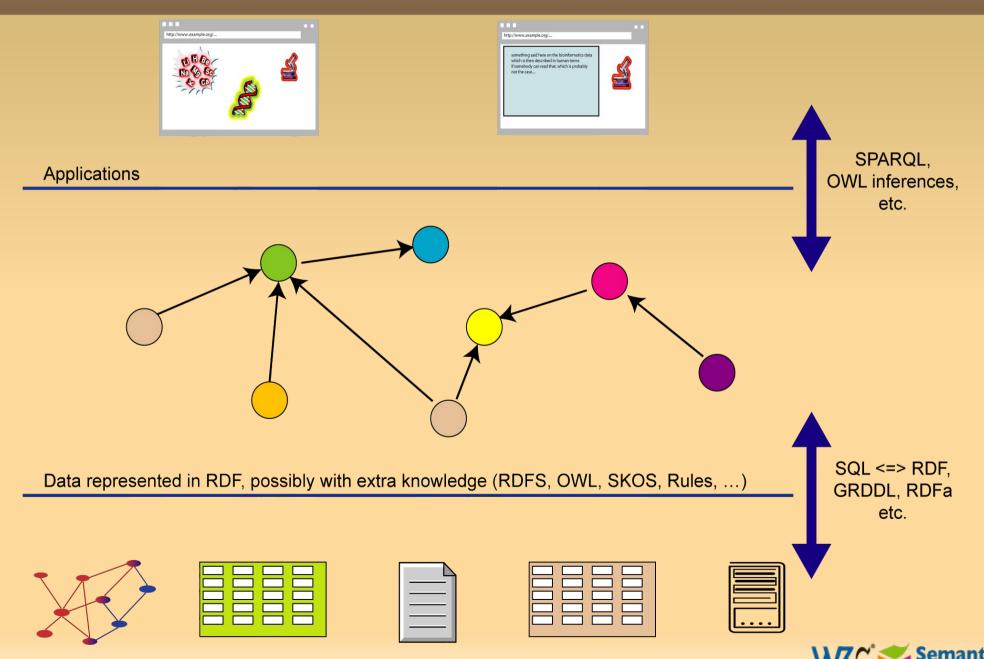








So where is the Semantic Web? (cont)



Data in various formats

Public datasets are accumulating

- IgentaConnect bibliographic metadata storage: over 200 million triplets
- RDFS/OWL Representation of WordNet: also downloadable as 150MB of RDF/XML
- "Département/canton/commune" structure of France published by the French Statistical Institute
- Geonames Ontology and Data: 6 million geographical features
- "dbpedia": infobox data of Wikipedia into RDF
- Note the "Billion Triple Challenge 2009"!



The network effect

- The usage of URI-s mean that we can link any data to any data on the Web
- The "network effect" is extended to the data on the Web
- "Mashup on steroids" become possible



Practical applications

Comes later today...



In the end...



- More an more data should be "published" on the Web
 - this can lead to the "network effect" on data
- New breeds of applications come to the fore
 - "mashups on steroids"
 - better representation and usage of community knowledge
 - new customization possibilities
 - ...

