WORLD WIDE WEB consortium
Australian Office

Introduction and Applications of the Semantic Web

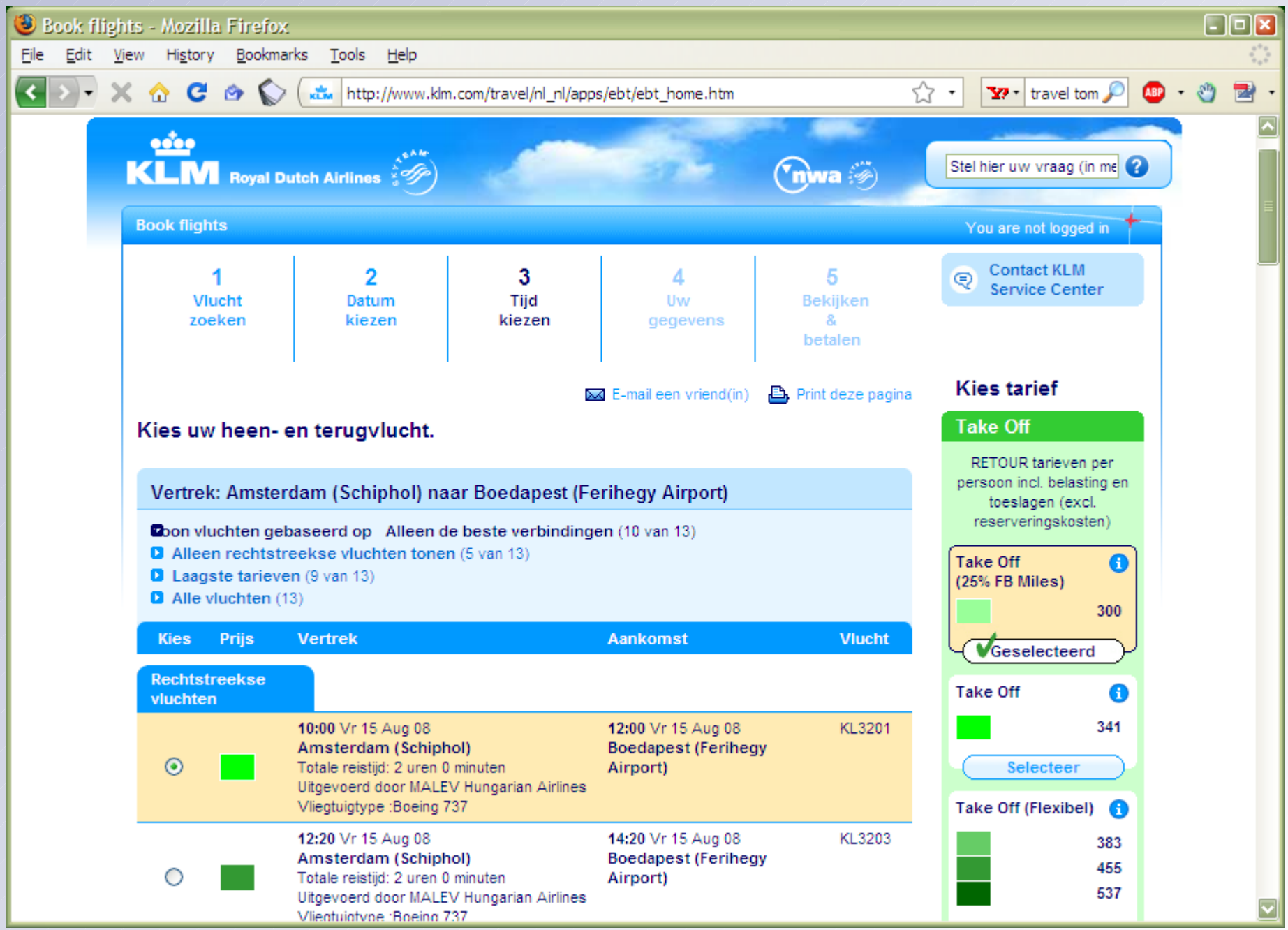
Slides by: Ivan Herman, W3C

http://www.w3.org/2009/Talks/05-Oz-IntroSW-IH/
http://www.w3.org/2009/Talks/0615-SanJose-tutorial-IH/

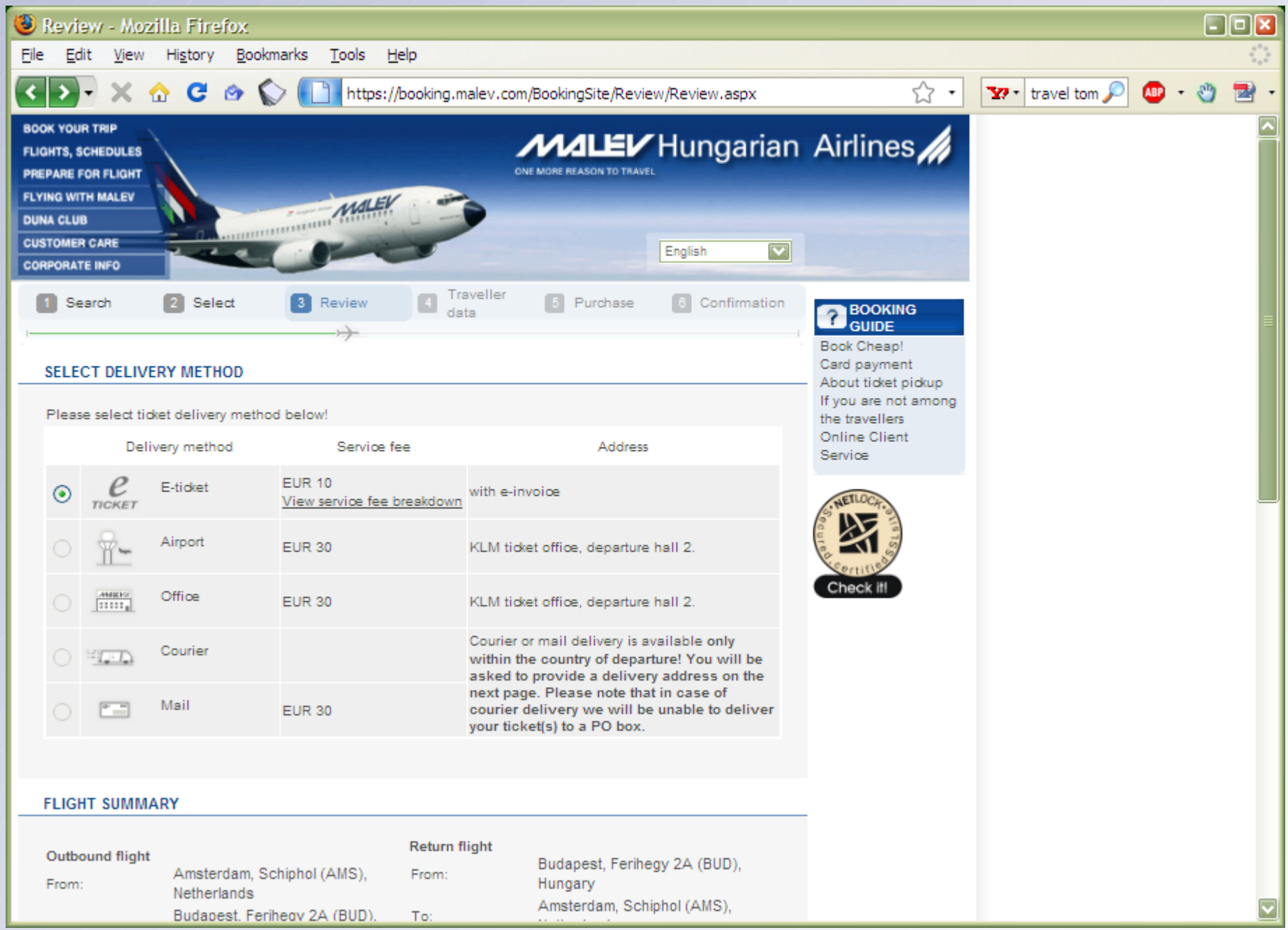
Let's organize a trip to Budapest from Amsterdam using the Web!

You try to find a proper flight with …

… a big, reputable airline, or …

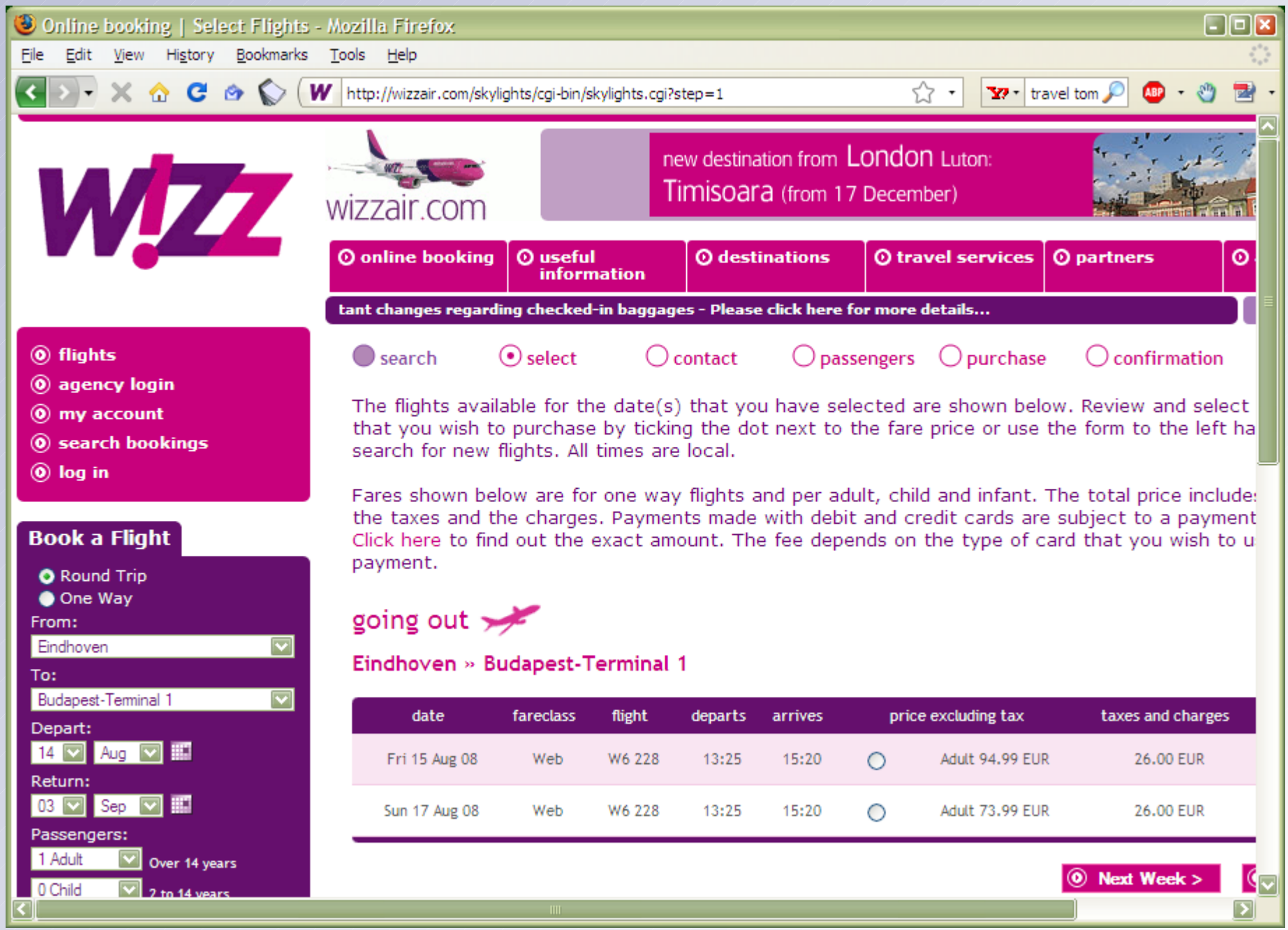


… the airline of the target country, or …



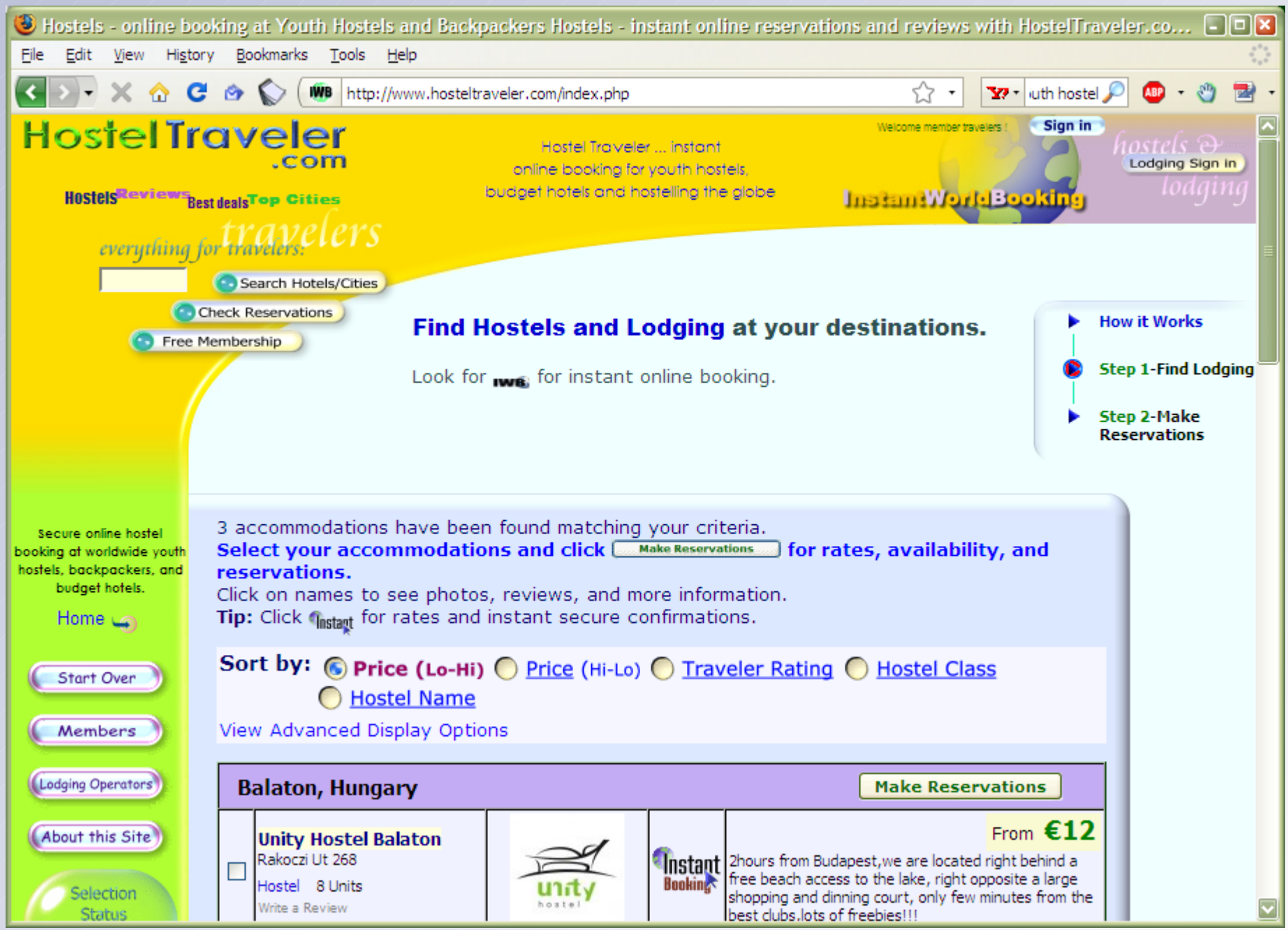
W3C Semantic Web

… or a low cost one



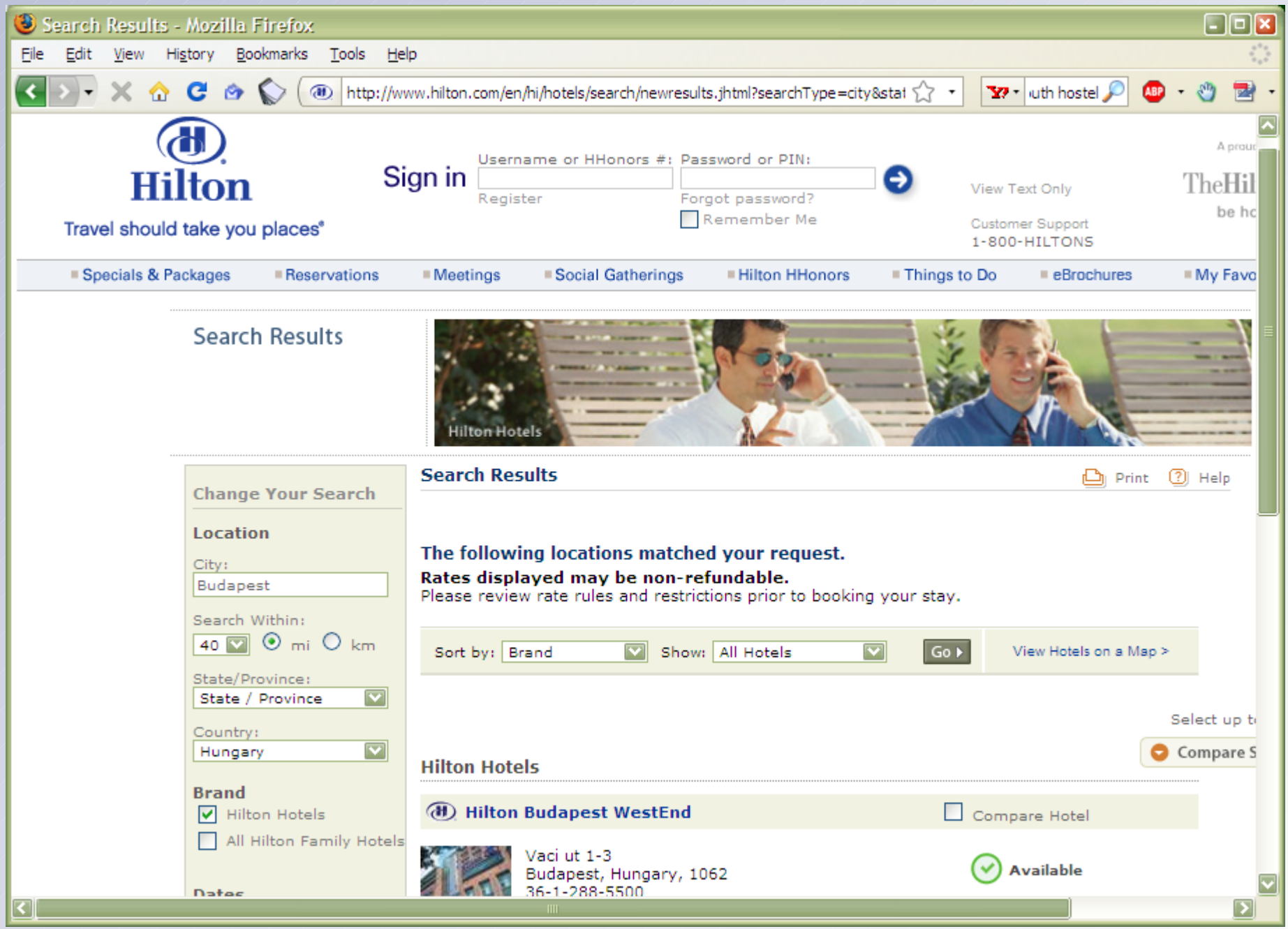
You have to find a hotel, so you look for…

… a really cheap accommodation, or …

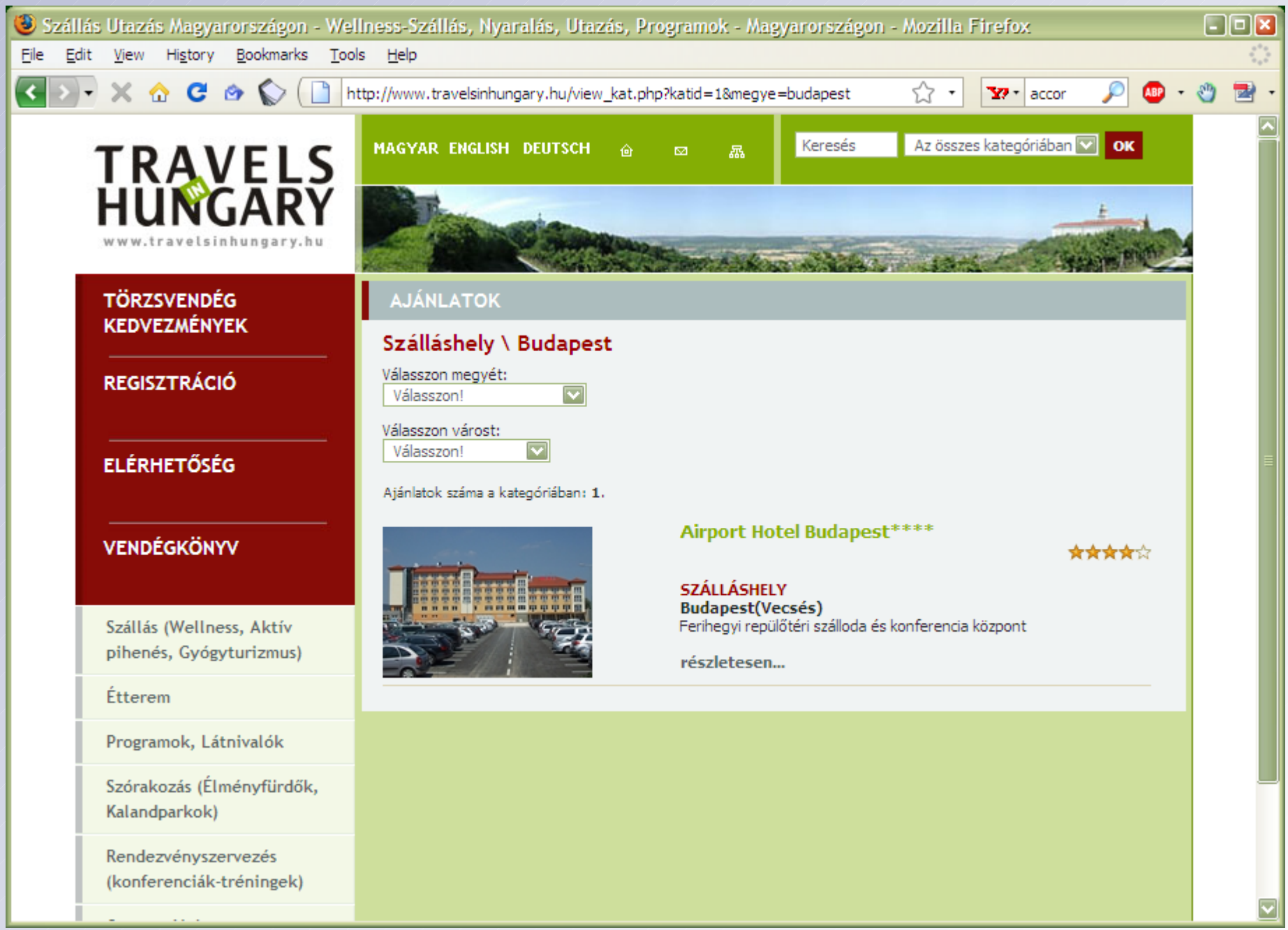


W3C Semantic Web

… or a really luxurious one, or …

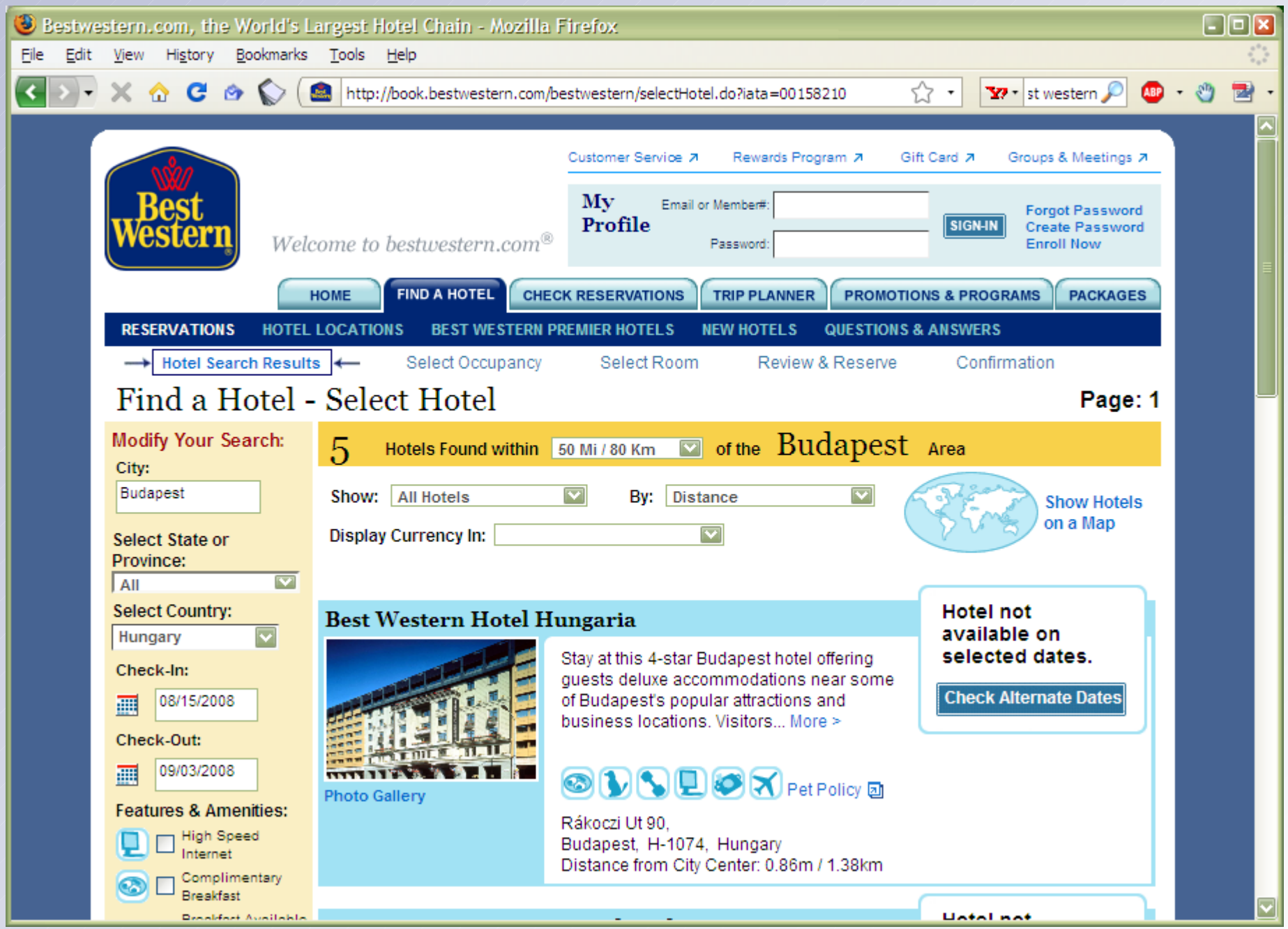


… an intermediate one …



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

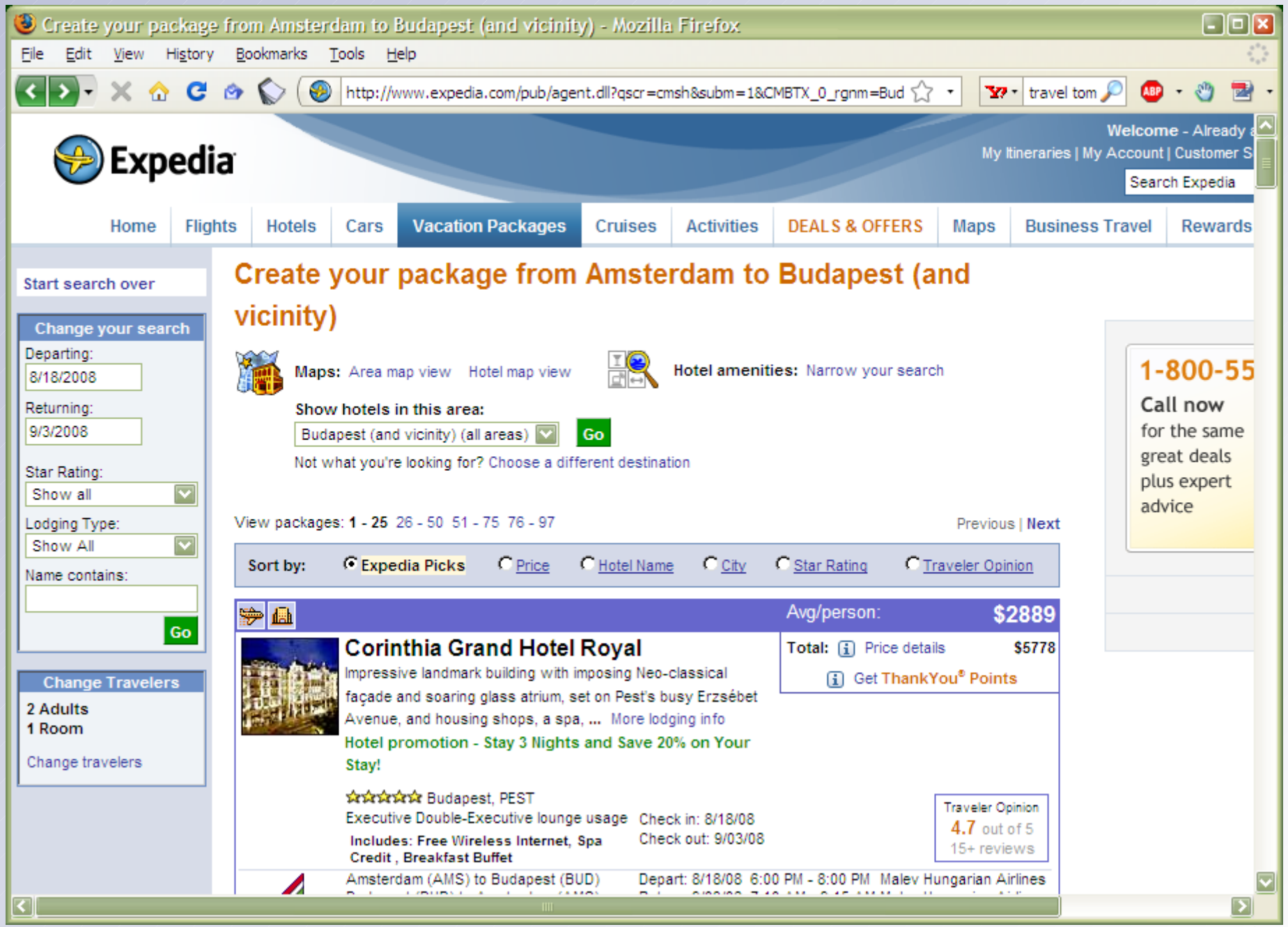
… this one could work



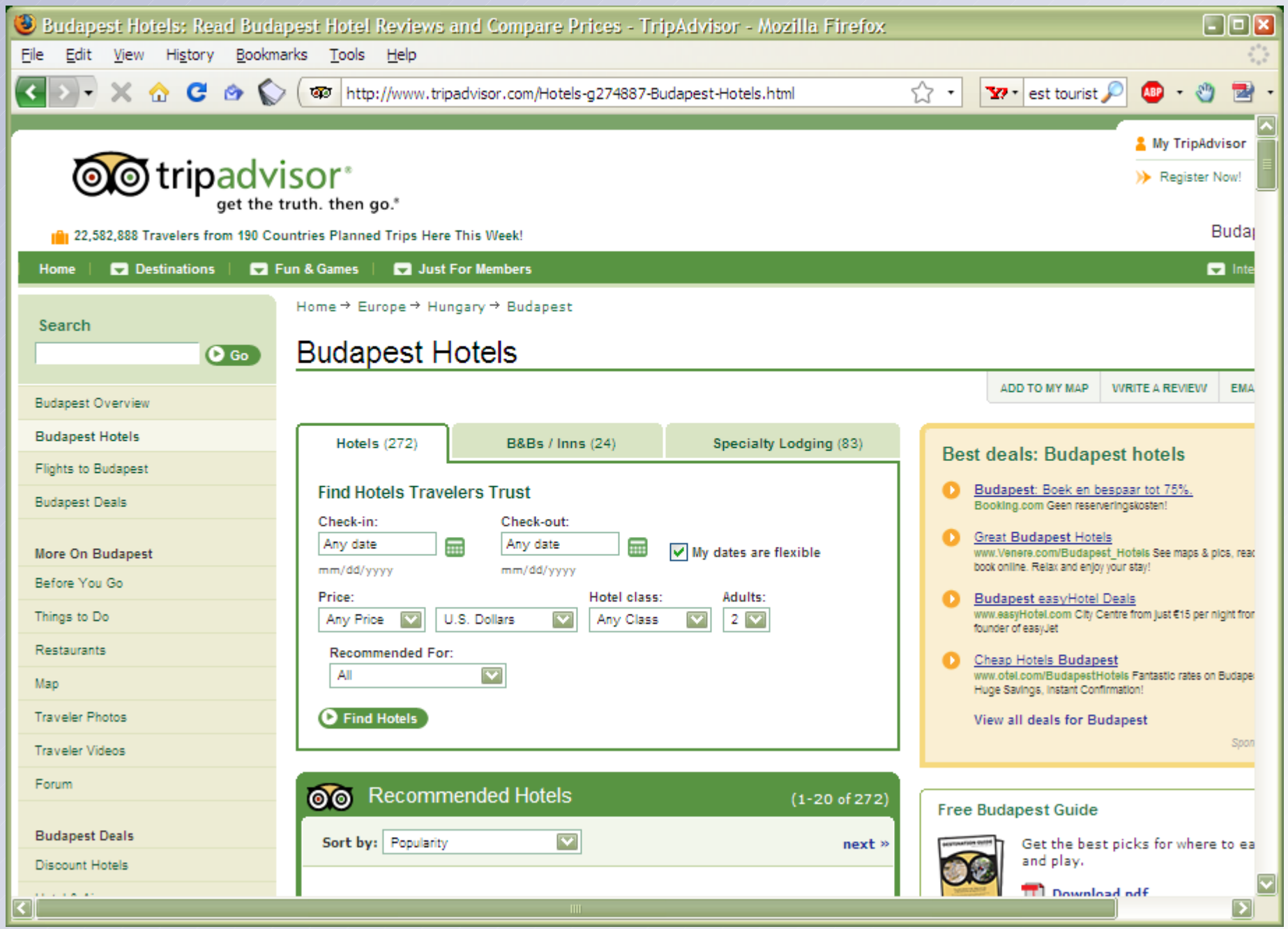
W3C Semantic Web

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

… like this one, or…

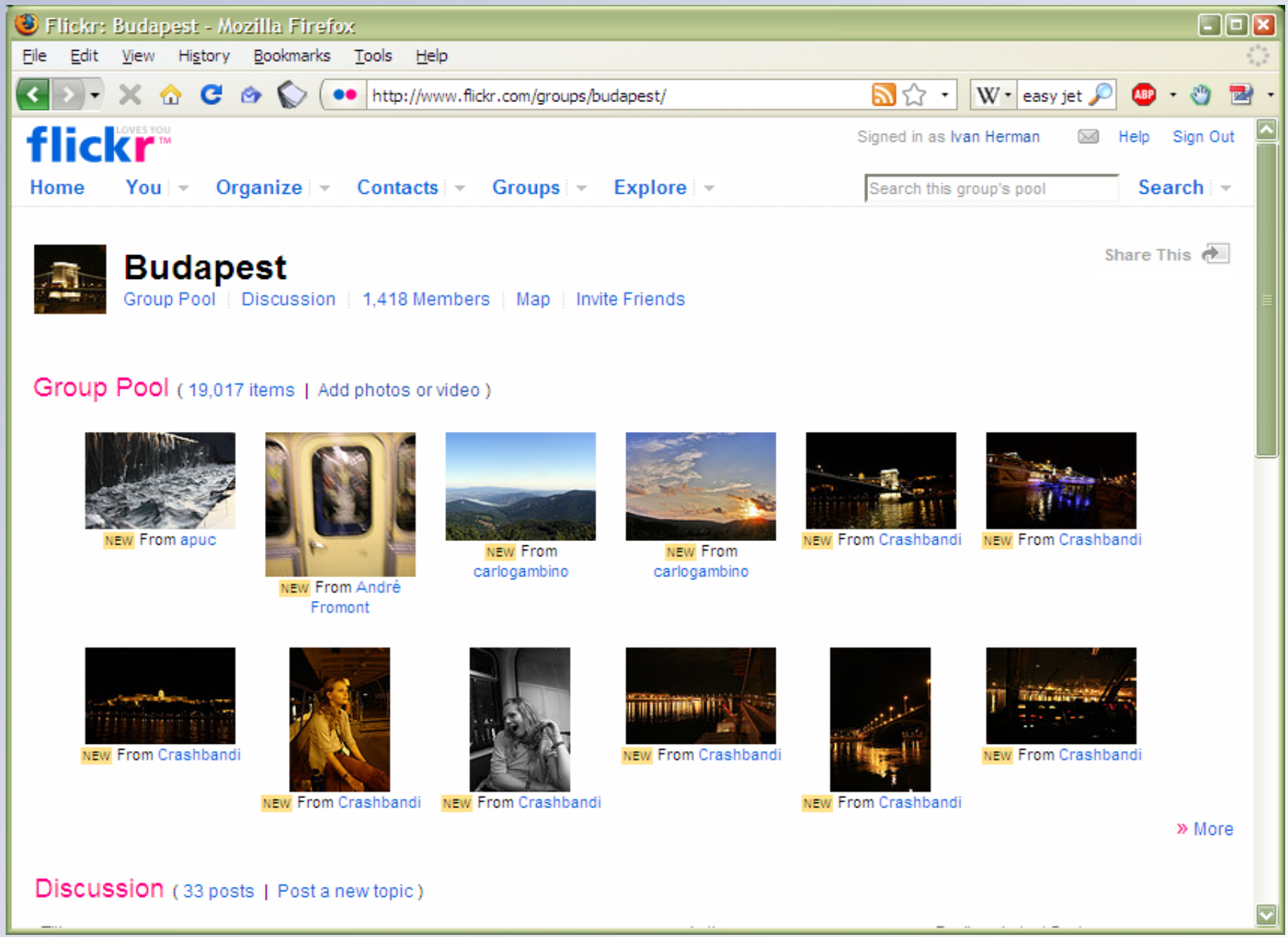


… or this one

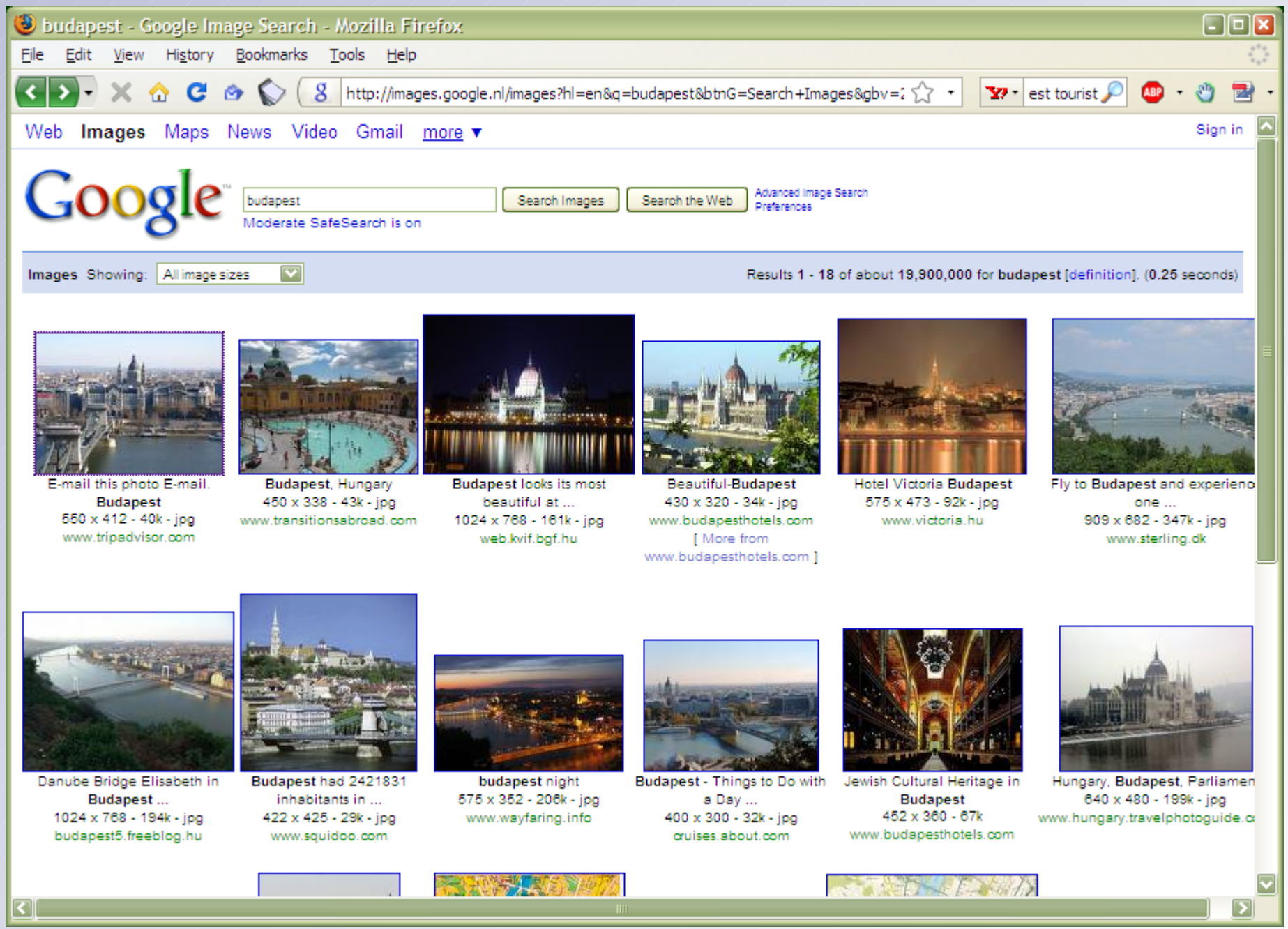

Budapest Hotels: Read Budapest Hotel Reviews and Compare Prices - TripAdvisor - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.tripadvisor.com/Hotels-g274887-Budapest-Hotels.html
tripadvisor
get the truth. then go.
My TripAdvisor
Register Now!
22,582,888 Travelers from 190 Countries Planned Trips Here This Week!
Home Destinations Fun & Games Just For Members
Search
Go
Budapest Overview
Budapest Hotels
Flights to Budapest
Budapest Deals
More On Budapest
Before You Go
Things to Do
Restaurants
Map
Traveler Photos
Traveler Videos
Forum
Budapest Deals
Discount Hotels
Home → Europe → Hungary → Budapest
Budapest Hotels
ADD TO MY MAP
WRITE A REVIEW
Hotels (272)
B&Bs / Inns (24)
Specialty Lodging (83)
Find Hotels Travelers Trust
Check-in:
Any date
mm/dd/yyyy
Check-out:
Any date
mm/dd/yyyy
My dates are flexible
Price:
Any Price
U.S. Dollars
Hotel class:
Any Class
Adults:
2
Recommended For:
All
Find Hotels
Best deals: Budapest hotels
Budapest: Boek en bespaar tot 75%.
Booking.com Geen reserveringskosten!
Great Budapest Hotels
Budapest easyHotel Deals
Cheap Hotels Budapest
View all deals for Budapest
Recommended Hotels
(1-20 of 272)
Sort by: Popularity
next »
Free Budapest Guide
Get the best picks for where to ea
and play.
Download pdf

You may want to know something about Budapest; look for some photo-graphs…

… on flickr …

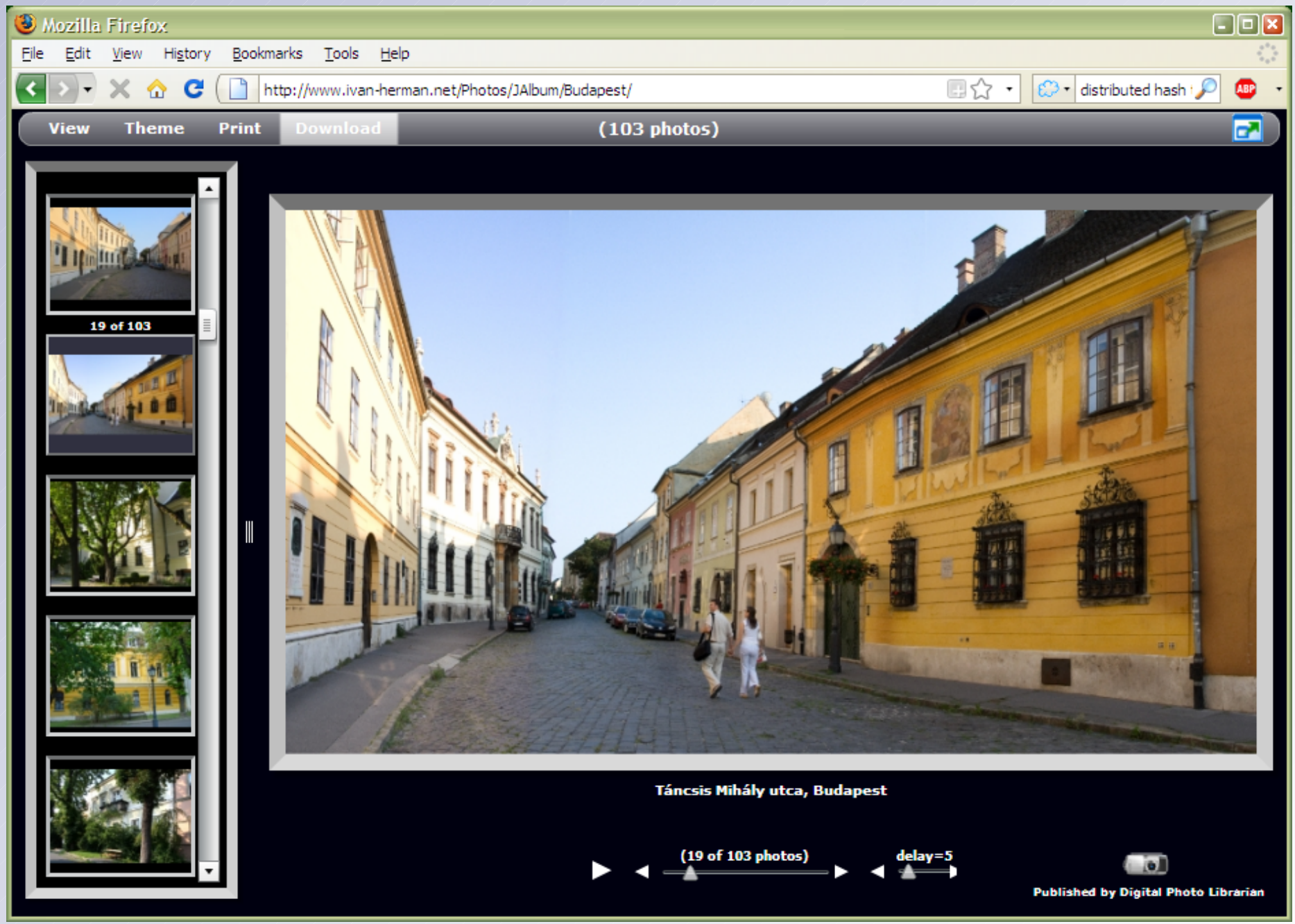

Flickr: Budapest - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.flickr.com/groups/budapest/
easy jet
flickr
Signed in as Ivan Herman
Help
Sign Out
Home
You
Organize
Contacts
Groups
Explore
Search this group's pool
Search
Budapest
Share This
Group Pool
Discussion
1,418 Members
Map
Invite Friends
Group Pool (19,017 items | Add photos or video)
NEW From apuc
NEW From André Fromont
NEW From carlogambino
NEW From carlogambino
NEW From Crashbandi
NEW From Crashbandi
NEW From Crashbandi
NEW From Crashbandi
NEW From Crashbandi
NEW From Crashbandi
NEW From Crashbandi
» More
Discussion (33 posts | Post a new topic)

… on Google …

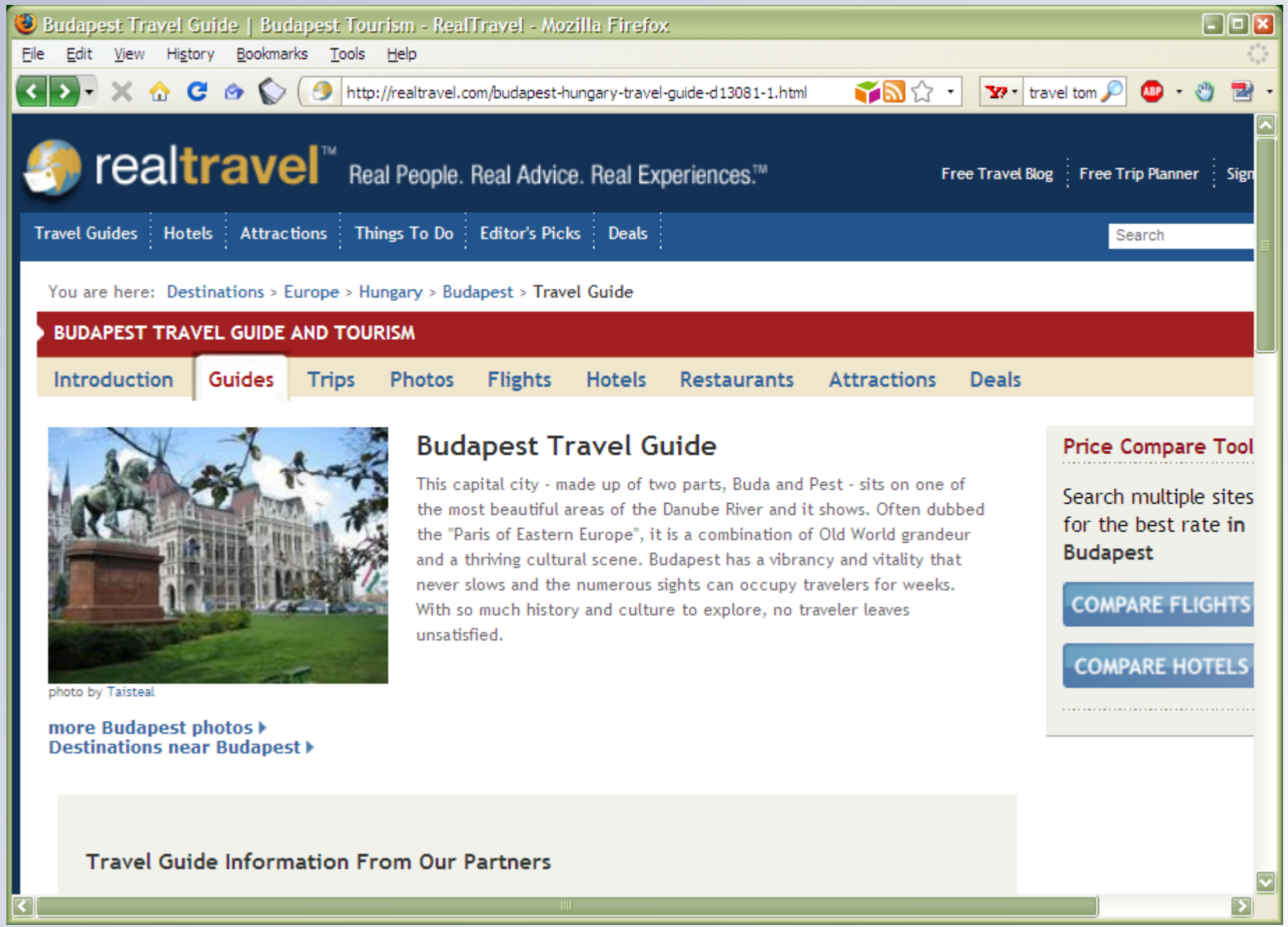


W3C Semantic Web

… or you can look at mine



...or at a (social) travel site



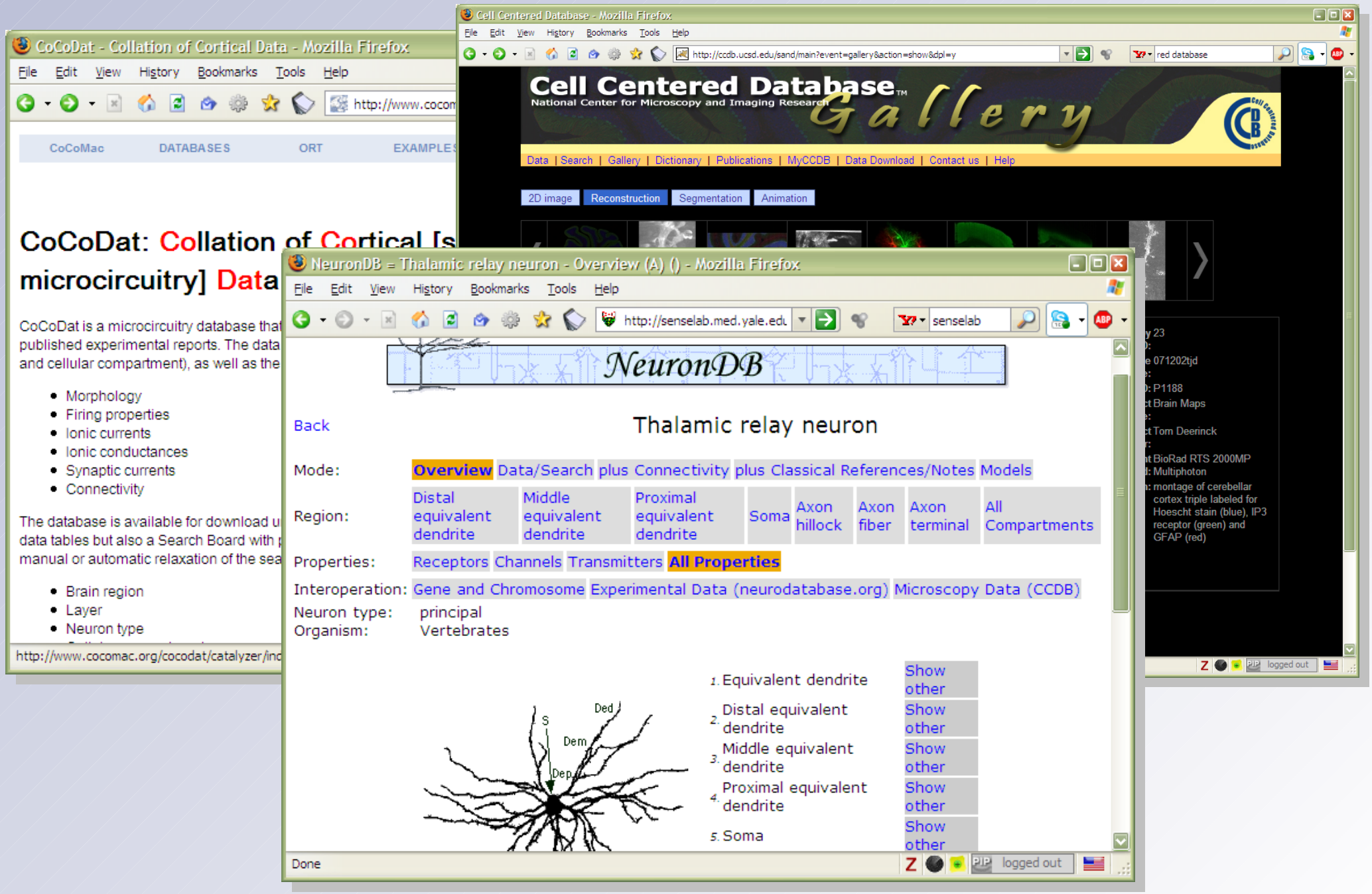
What happened here?

- 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

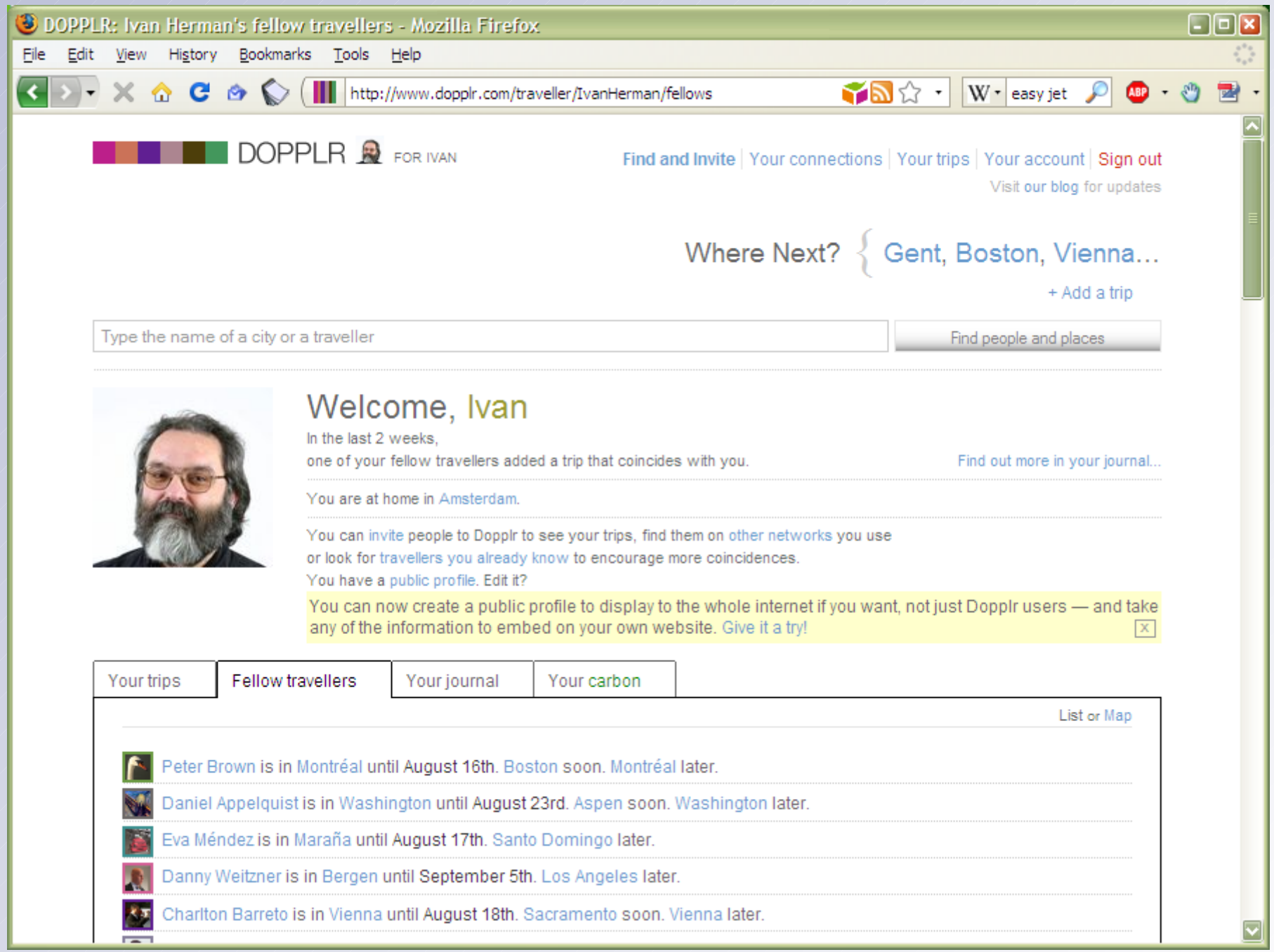
- Specialized sites (Expedia, TripAdvisor) do a bit more:
  - they gather and combine data from other sources (usually with the approval of the data owners)
  - but they still control how you see those sources
- But sometimes you want to personalize: access the original data and combine it yourself!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

Here is another example…

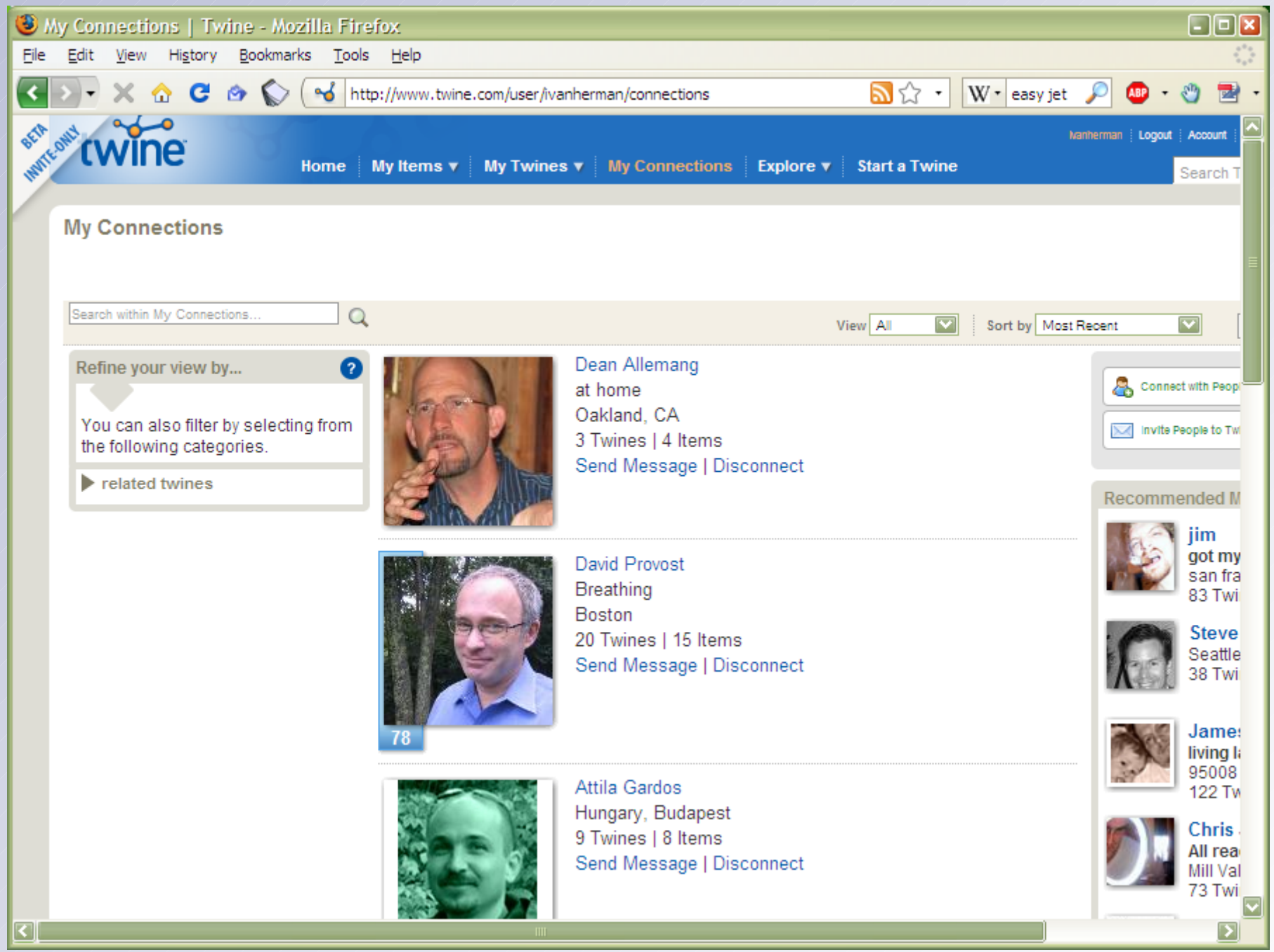


Another example: social sites. I have a list of "friends" by…

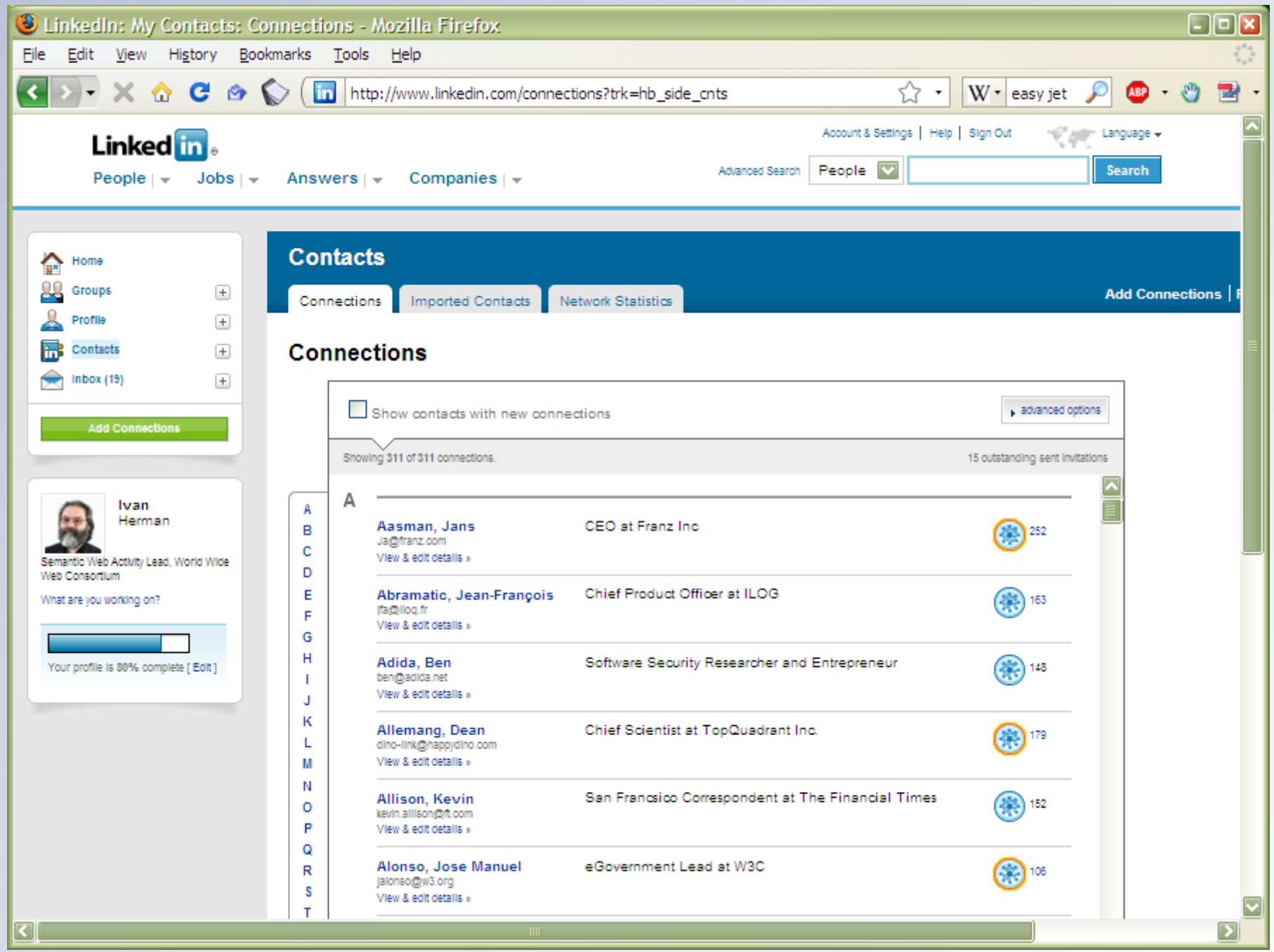
… Dopplr,


DOPPLR: Ivan Herman's fellow travellers - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.dopplr.com/traveller/IvanHerman/fellows
easy jet
DOPPLR FOR IVAN
Find and Invite | Your connections | Your trips | Your account | Sign out
Visit our blog for updates
Where Next? { Gent, Boston, Vienna…
+ Add a trip
Type the name of a city or a traveller
Find people and places
Welcome, Ivan
In the last 2 weeks,
one of your fellow travellers added a trip that coincides with you.
Find out more in your journal...
You are at home in Amsterdam.
You can invite people to Dopplr to see your trips, find them on other networks you use
or look for travellers you already know to encourage more coincidences.
You have a public profile. Edit it?
You can now create a public profile to display to the whole internet if you want, not just Dopplr users — and take any of the information to embed on your own website. Give it a try!
Your trips
Fellow travellers
Your journal
Your carbon
List or Map
Peter Brown is in Montréal until August 16th. Boston soon. Montréal later.
Daniel Appelquist is in Washington until August 23rd. Aspen soon. Washington later.
Eva Méndez is in Maraña until August 17th. Santo Domingo later.
Danny Weitzner is in Bergen until September 5th. Los Angeles later.
Charlton Barreto is in Vienna until August 18th. Sacramento soon. Vienna later.

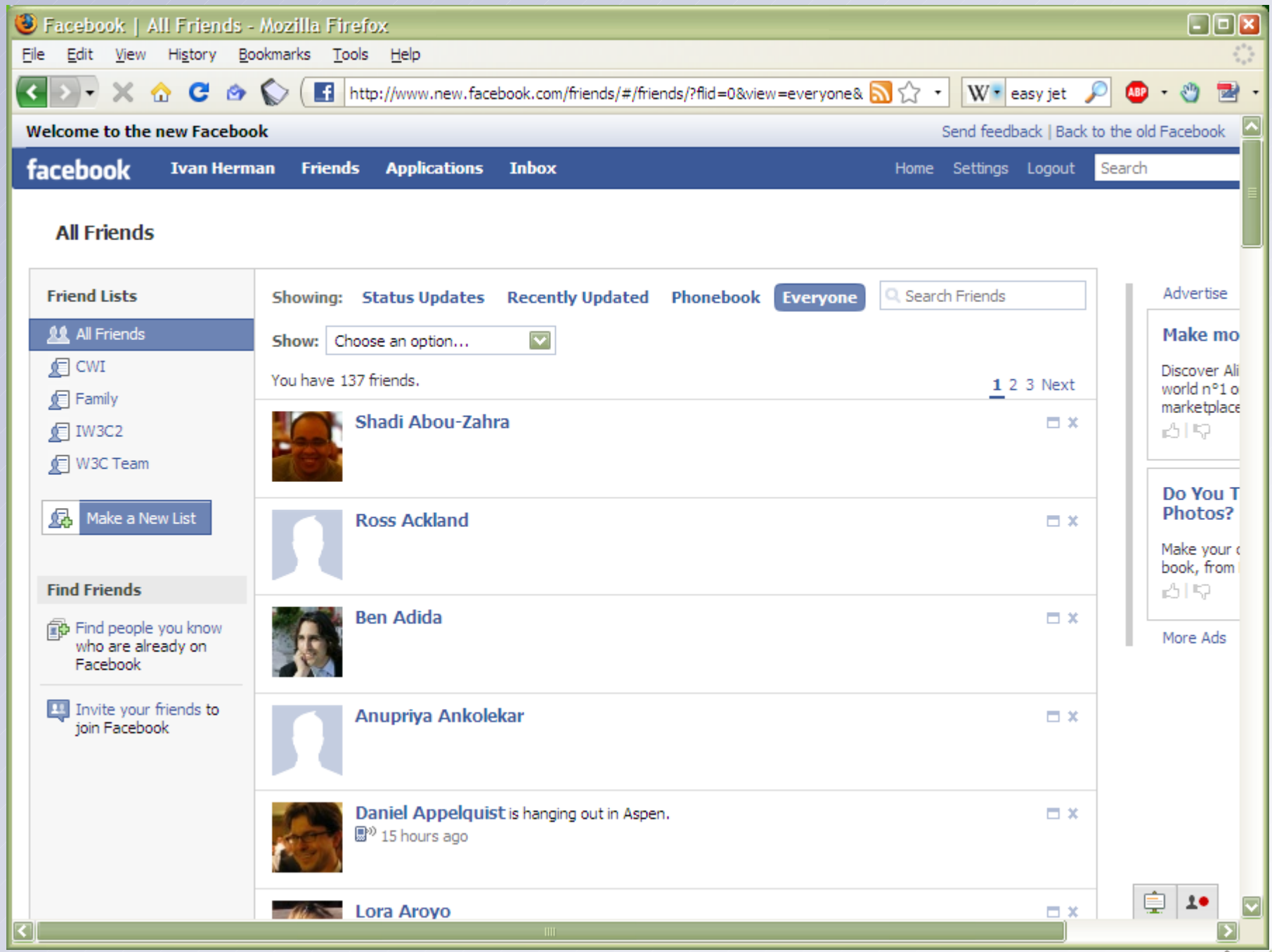
… Twine,


My Connections | Twine - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.twine.com/user/ivanherman/connections
easy jet
BETA INVITE-ONLY
twine
Home
My Items
My Twines
My Connections
Explore
Start a Twine
Logout
Account
My Connections
Search within My Connections...
View All
Sort by Most Recent
Refine your view by...
You can also filter by selecting from the following categories.
related twines
Dean Allemang
at home
Oakland, CA
3 Twines | 4 Items
Send Message | Disconnect
David Provost
Breathing
Boston
20 Twines | 15 Items
Send Message | Disconnect
78
Attila Gardos
Hungary, Budapest
9 Twines | 8 Items
Send Message | Disconnect
Connect with Peop
Invite People to Tw
Recommended M
jim
got my
san fra
83 Twi
Steve
Seattle
38 Twi
Jame
living l
95008
122 Tw
Chris
All rea
Mill Val
73 Twi

… LinkedIn,


LinkedIn: My Contacts: Connections - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.linkedin.com/connections?trk=hb_side_cnts
easy jet
LinkedIn
Account & Settings | Help | Sign Out
Language
People | Jobs | Answers | Companies |
Advanced Search
People
Search
Home
Groups
Profile
Contacts
Inbox (19)
Add Connections
Ivan Herman
Semantic Web Activity Lead, World Wide Web Consortium
What are you working on?
Your profile is 80% complete [Edit]
Contacts
Connections
Imported Contacts
Network Statistics
Add Connections
Connections
Show contacts with new connections
advanced options
Showing 311 of 311 connections.
15 outstanding sent invitations
A
Aasman, Jans
Ja@franz.com
View & edit details »
CEO at Franz Inc
252
Abramatic, Jean-François
jfa@ilog.fr
View & edit details »
Chief Product Officer at ILOG
163
Adida, Ben
ben@adida.net
View & edit details »
Software Security Researcher and Entrepreneur
148
Allemang, Dean
dino-link@happydino.com
View & edit details »
Chief Scientist at TopQuadrant Inc.
179
Allison, Kevin
kevin.allison@ft.com
View & edit details »
San Francisco Correspondent at The Financial Times
152
Alonso, Jose Manuel
jalonso@w3.org
View & edit details »
eGovernment Lead at W3C
106
A B C D E F G H I J K L M N O P Q R S T

… and, of course, 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?

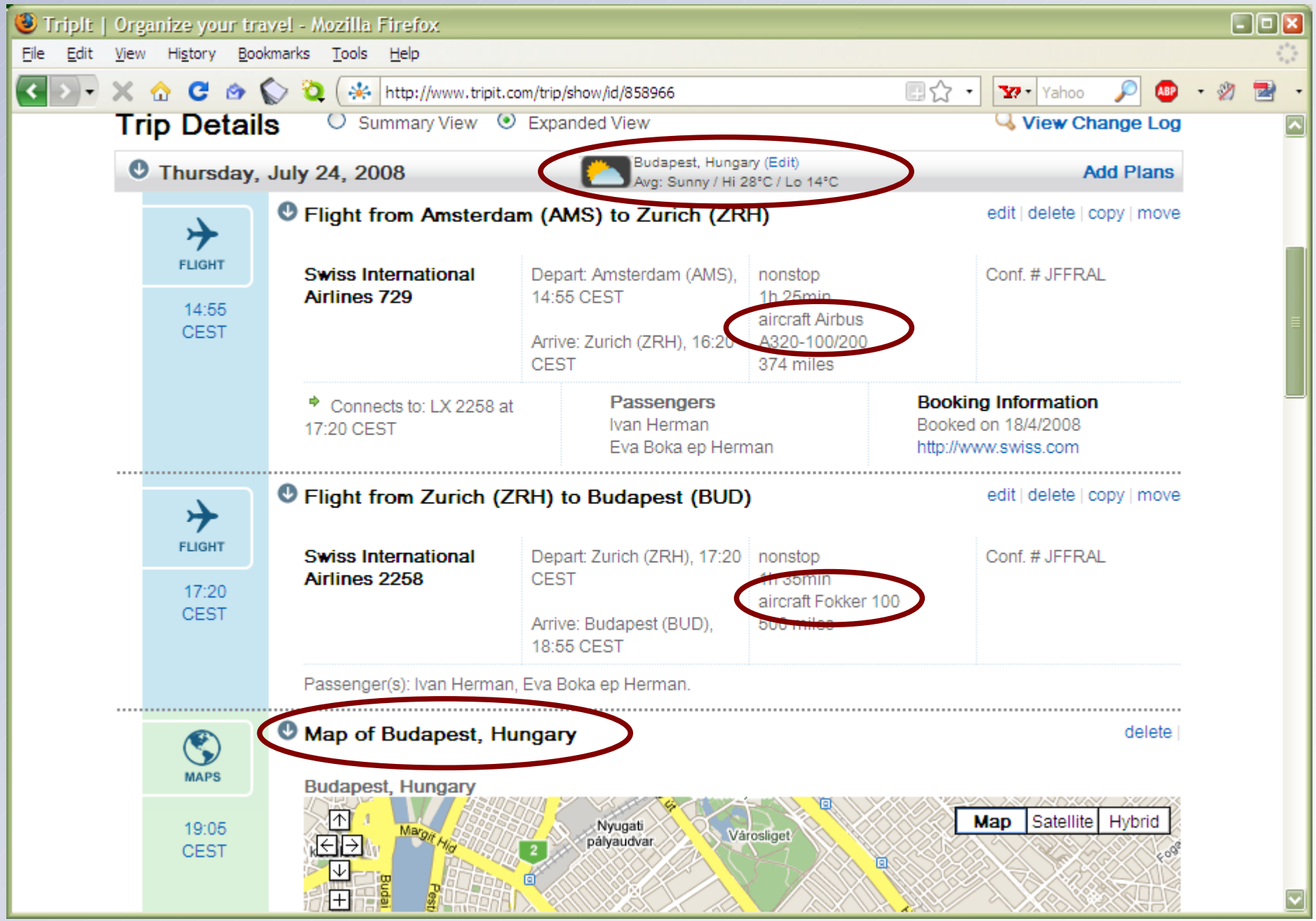
- Use the data on the Web the same way as we do with documents:
  - be able to link to data (independently of their presentation)
  - use that data the way I want (present it, mine it, etc)
  - agents, programs, scripts, etc, should be able to *interpret* part of that data

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 *standard* 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

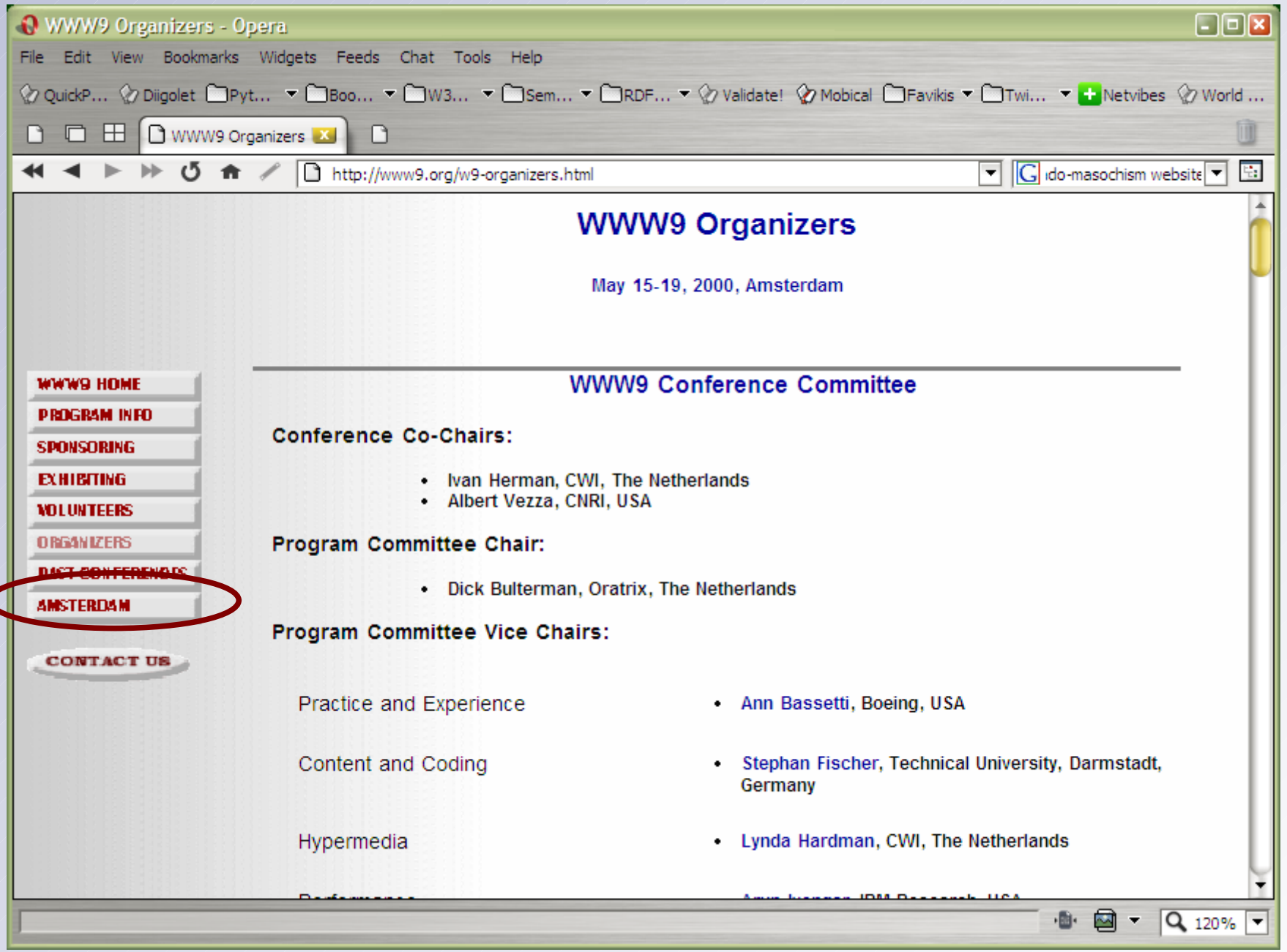
Steven's site on Amsterdam (done for some visiting friends)

W3C Semantic Web

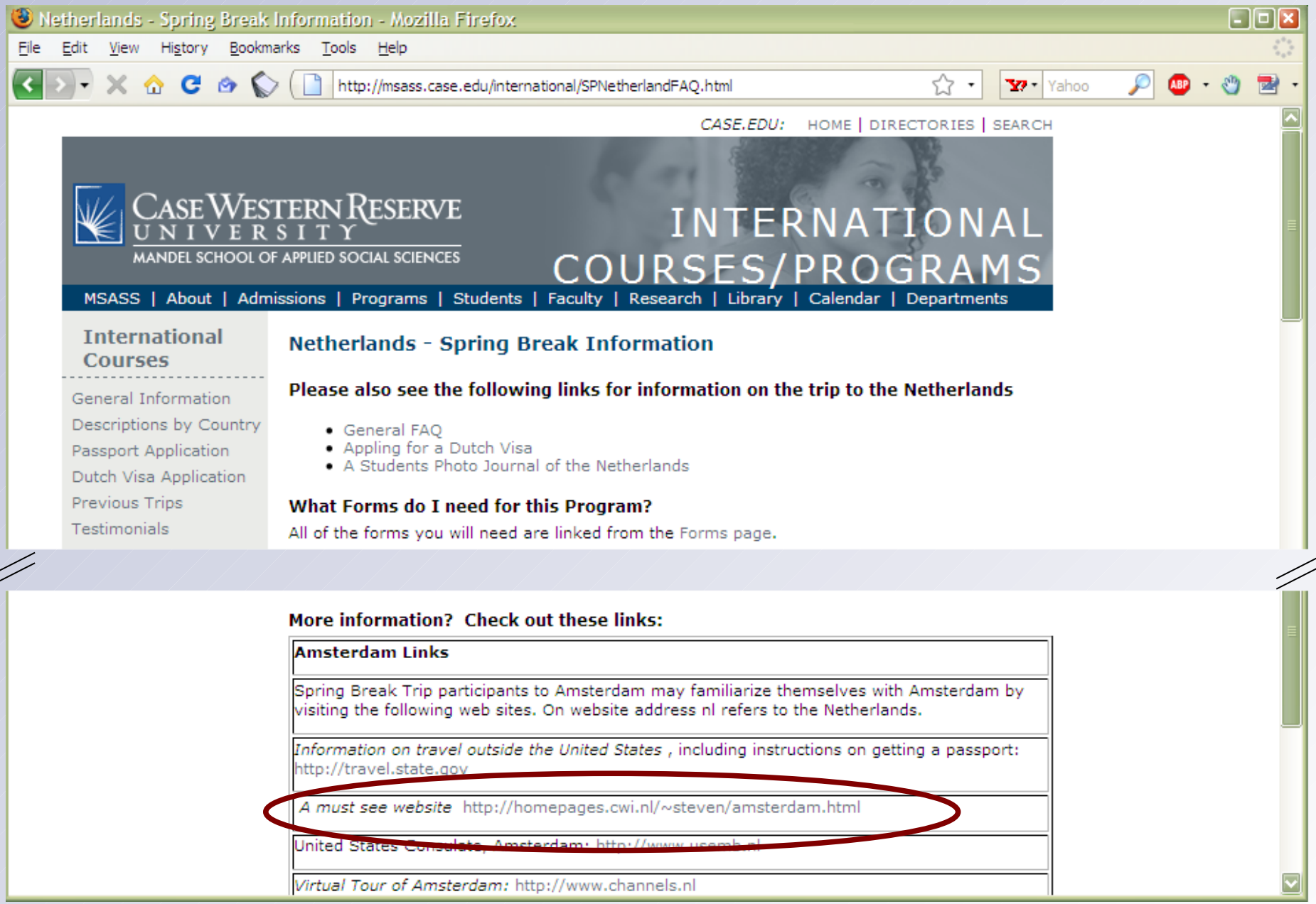
Then some magic happens…

- 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 *even if the author did not expect it…*

This could be expected…



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

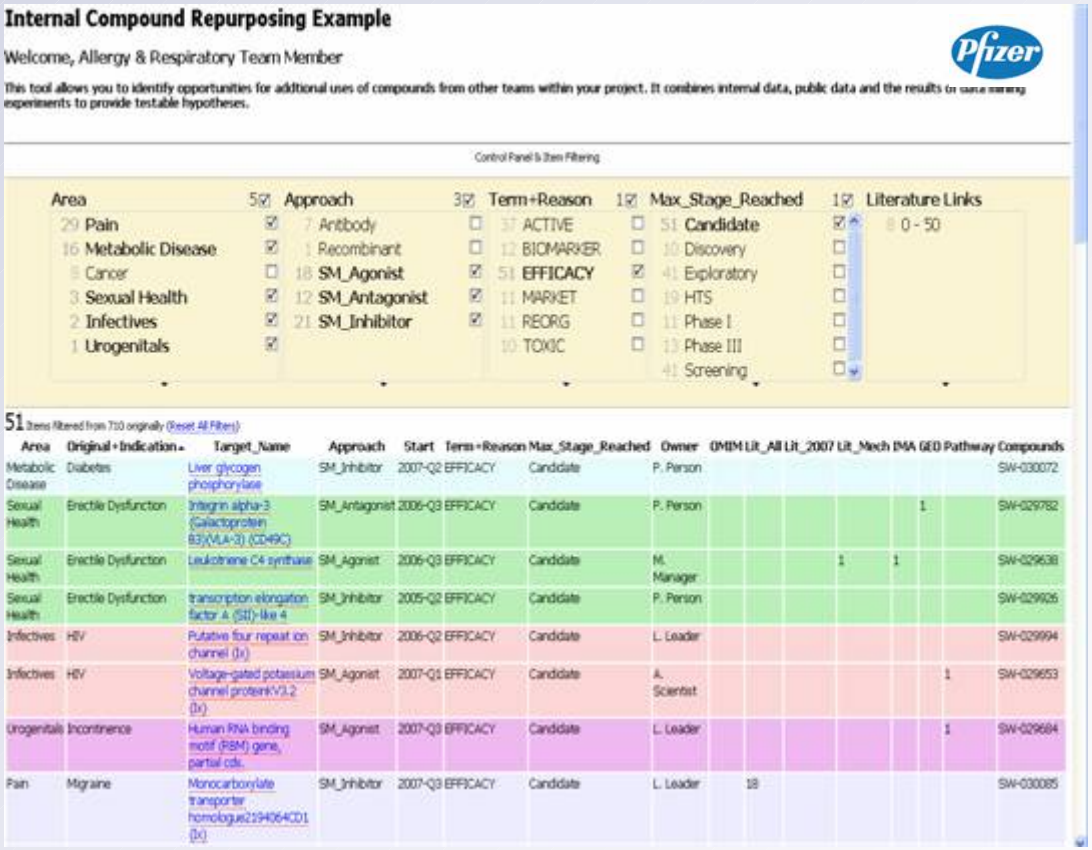


What would that mean for a Web of Data?

- 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



Courtesy of Nigel Wilkinson, Lee Harland, Pfizer Ltd, Melliyal Annamalai, Oracle (SWEO Case Study)

W3C Semantic Web

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:

Ivan Herman - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.w3.org/People/Ivan/

Ivan Herman

My Work at W3C | Contact information | Short CV | Upcoming trips | Public presentations

My Work at W3C

I am Semantic Web Activity Lead; that is my main work at W3C. I am member of IW3C2 (International World Wide Web Conference Committee) (the committee coordinating the yearly WWW conference series), serving as a liaison for W3C, and of SWSA (Semantic Web Science Association), the committee responsible for the International Semantic Web Conferences series.

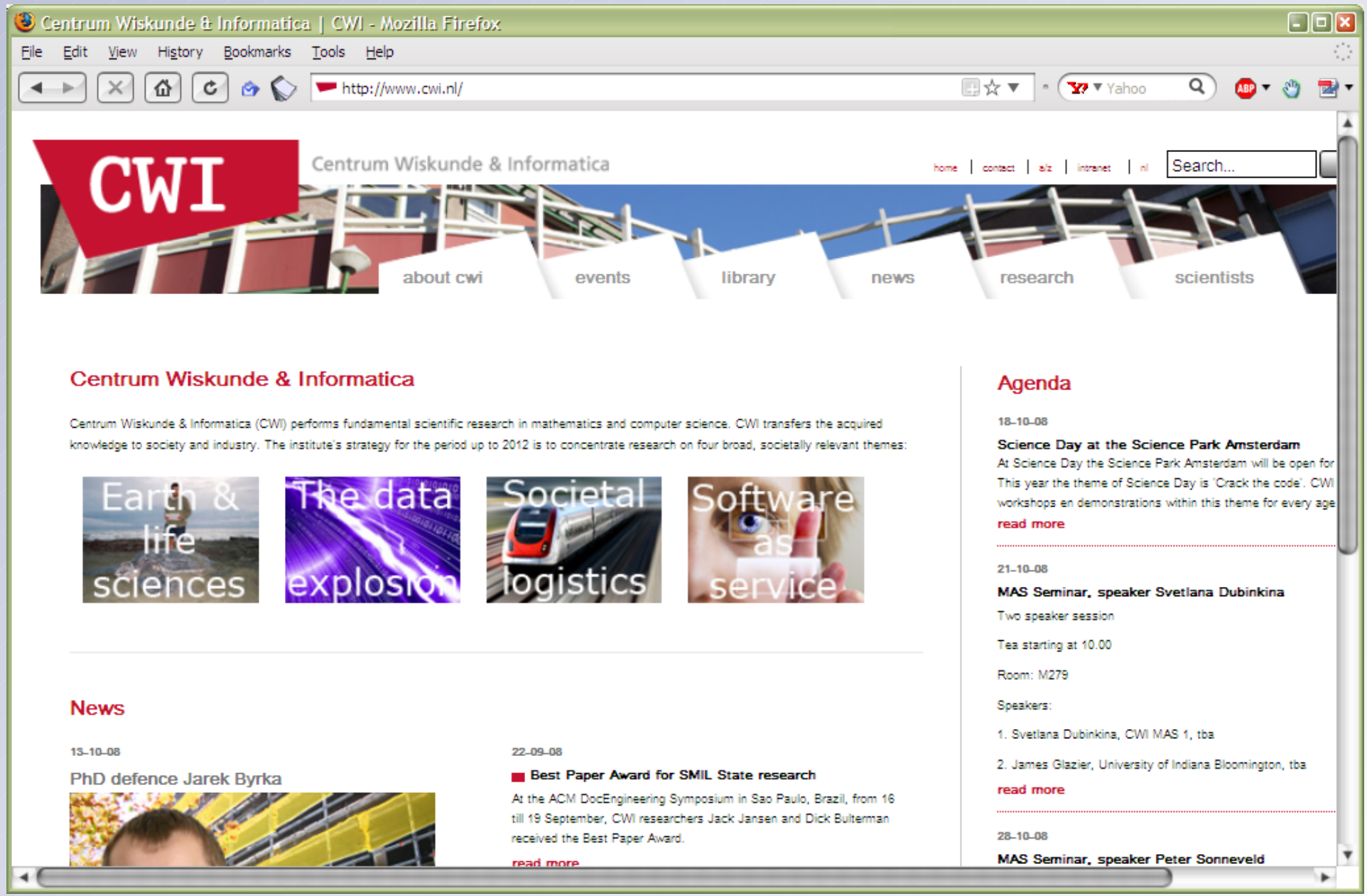
As part of my work, I also participate in lots of outreach activities, and I regularly make presentations, tutorials, etc. You can consult my list of presentations for further details.

Contact information

Email:
ivan@w3.org
(sha1sum: 5ac8032d5f8012aa1775ca2f03e1676bafd5e80b)
Postal address:
C/o Centre for Mathematics and Computer Sciences (CWI)
Kruislaan 413, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands.
Phone numbers:
phone: +31-20-5924163
mobile phone: +31-641044153
fax: +31-20-5924312
PGP/GPG:
My GnuPGP key and signature is available on-line.
FOAF:
You can either extract a short FOAF information from this page, of consult my more complete, public FOAF file.
Misc:

W3C Semantic Web

… leading to this page



- A human understands that this is an institution's home page
- He/she knows what it means (realizes that it is a research institute in the Netherlands)
- 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 an 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”
    - eg, if an address refers to Amsterdam, then this means it is also in the Netherlands

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

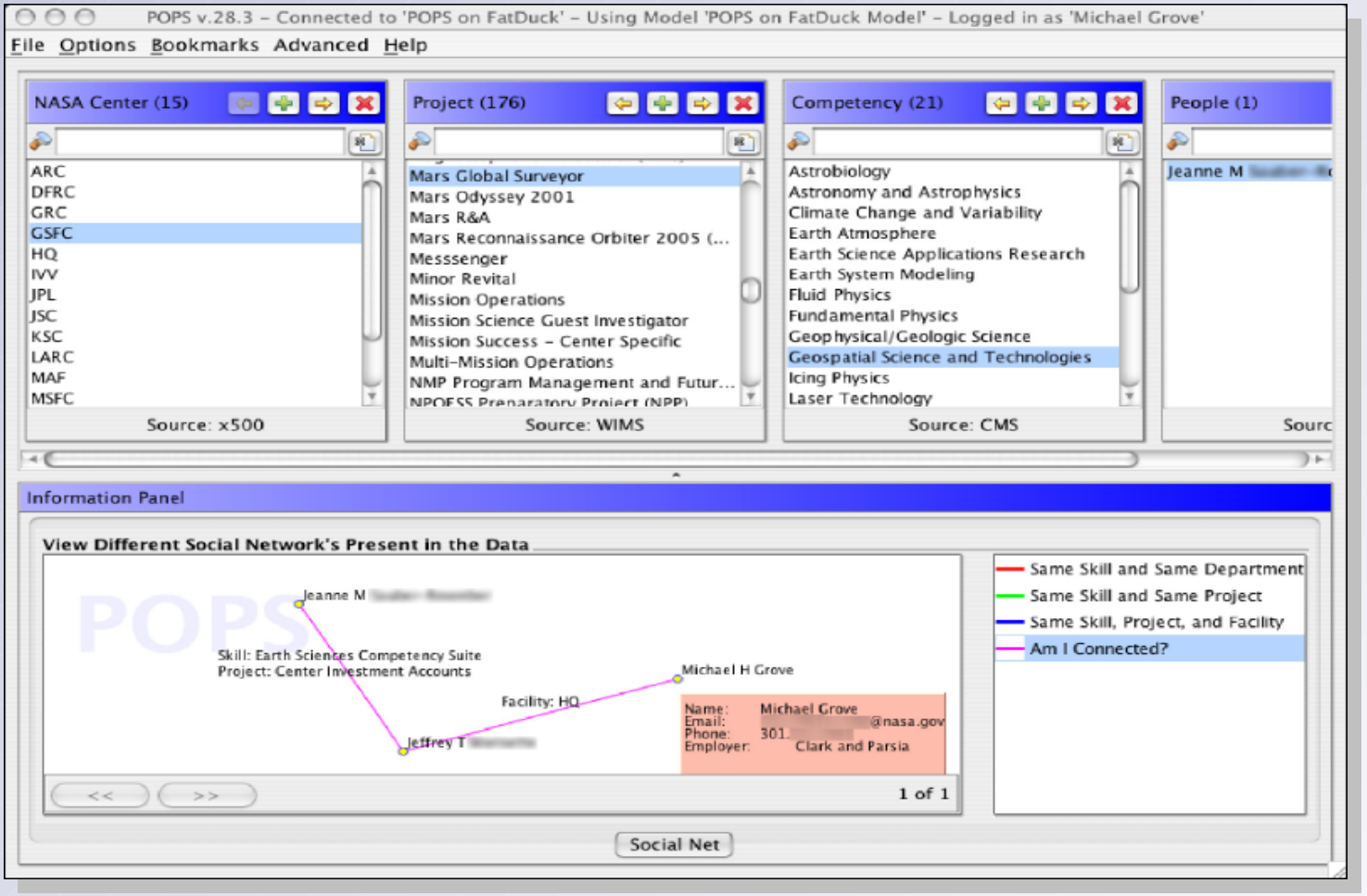
- It is that simple…
- Of course, the devil is in the details
  - a common model has to be provided for machines to describe, query, etc, the data and their connections
  - technologies should be around to “export” the data
  - the “classification” of the terms can become *very* complex for specific knowledge areas: this is where ontologies, thesauri, etc, enter the game…
  - but these details are fleshed out by experts as we speak!

Example: find the right experts at NASA

- NASA has nearly 70,000 civil servants over the whole of the US
- Their expertise is described in 6-7 databases, geographically distributed, with different data formats, access types…
- Task: find the right expert for a specific task within NASA!

Example: find the right experts at NASA

- Approach: integrate all the data with standard means, and describe the data and links using generic (and simple) vocabularies



Wait! Does it mean that I have to convert all my data in some way?

- Not necessarily; this would not always be feasible
- There are technologies to make your data *access-ible* to standard means without converting it
  - run-time “bridges” (eg, rewriting queries on the fly)
  - annotate existing data (eg, XHTML pages)
  - extract data from XHTML/XML files
  - etc
- Some of these techniques are still being developed

Example: "Linking Open Data Project"

- Goal: "expose" open datasets for integration
- *Set links among the data items* from different data-sets
- Set up query endpoints
- Altogether billions of relationships, millions of links…

Example data source: DBpedia

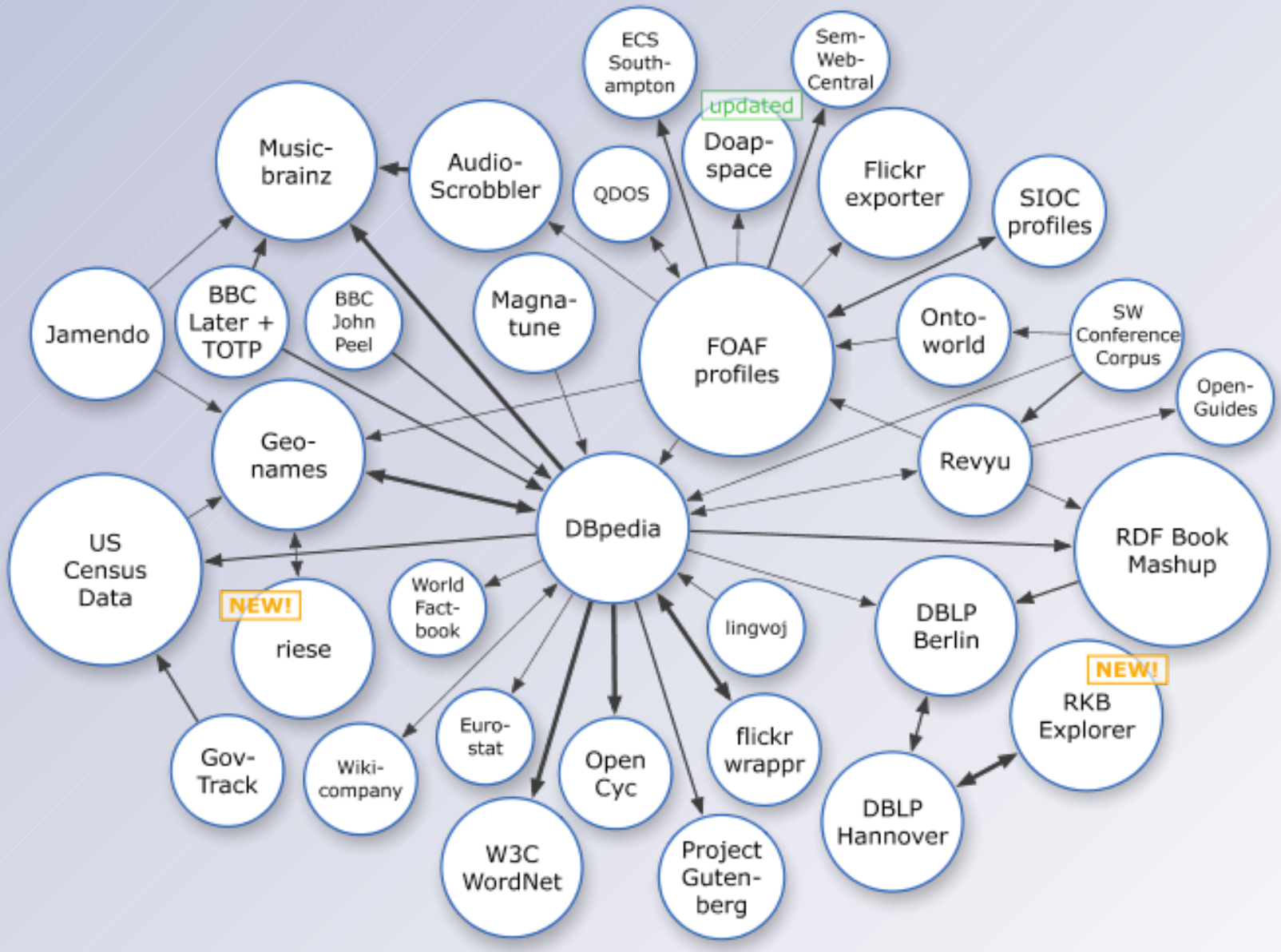
- DBpedia is a community effort to
  - extract structured (“infobox”) information from Wikipedia
  - provide a query endpoint to the dataset
  - interlink the DBpedia dataset with other datasets on the Web

Freie Universität Berlin

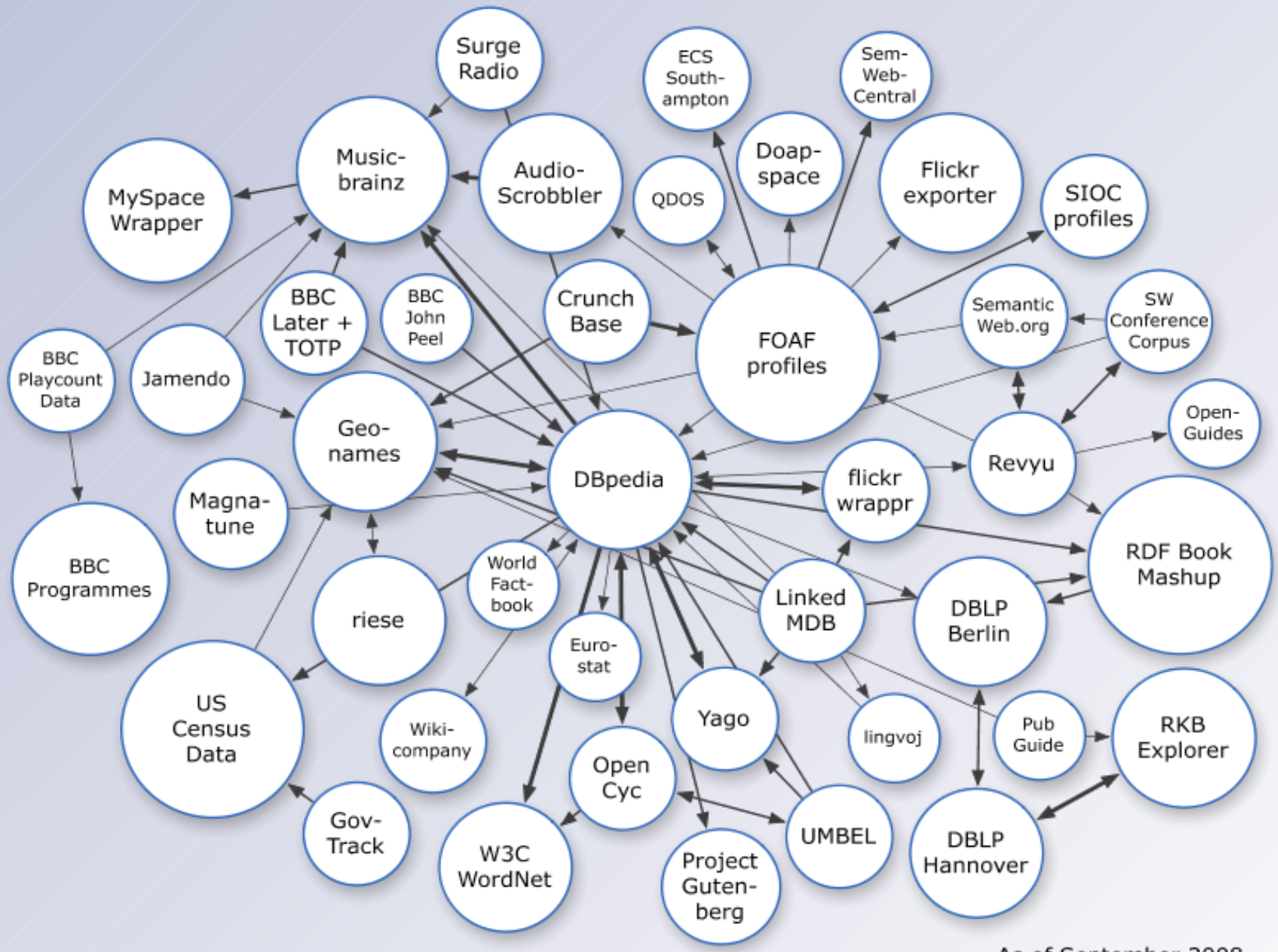
UNIVERSITÄT LEIPZIG

OPENLINK SOFTWARE

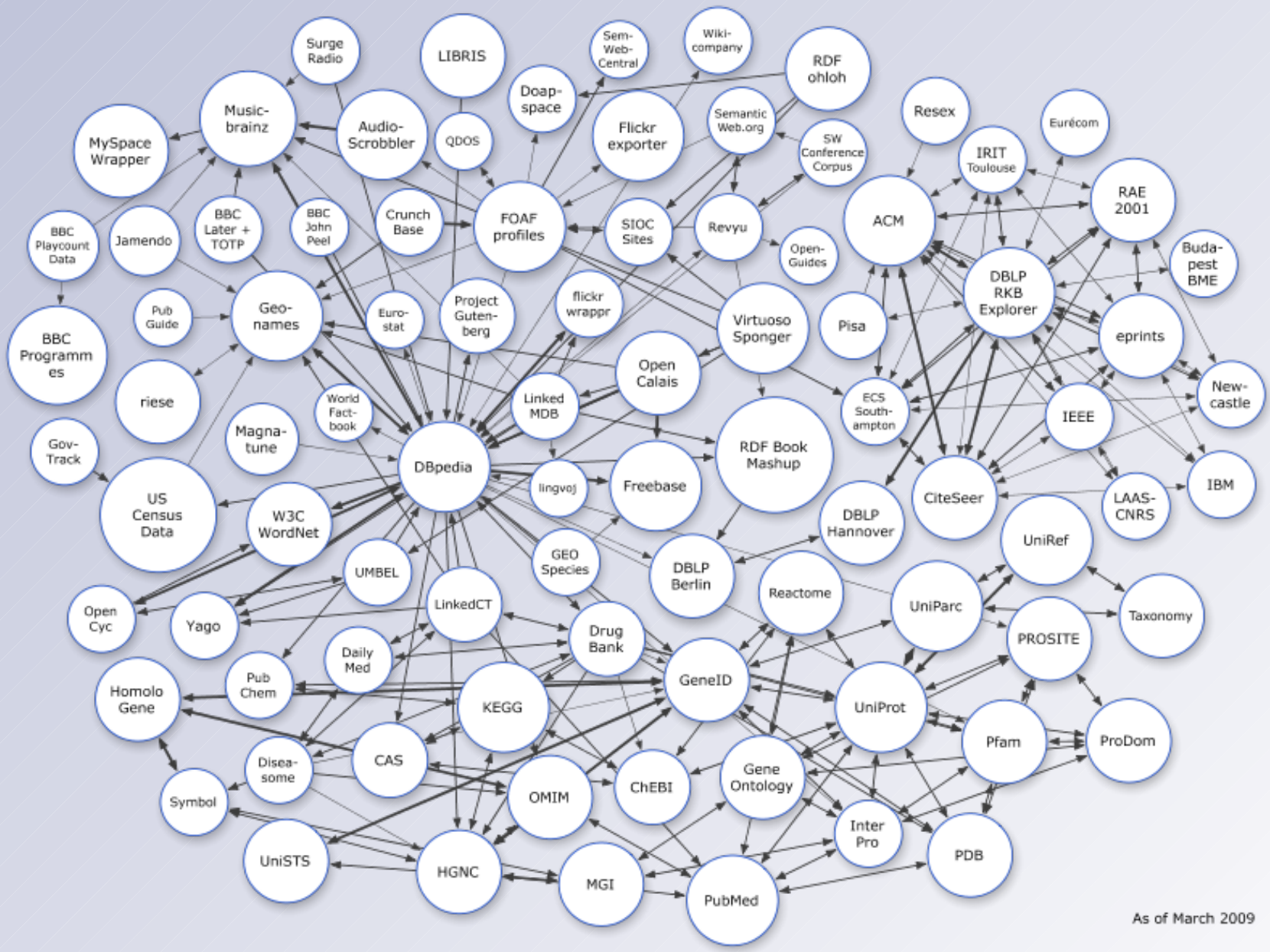
The LOD "cloud", March 2008



The LOD "cloud", September 2008



The LOD "cloud", March 2009



All this sounds nice, but isn't that just a dream?

The 2007 Gartner predictions

During the next 10 years, Web-based technologies will improve the ability to embed semantic structures [… it] will occur in multiple evolutionary steps…

By 2017, we expect the vision of the Semantic Web […] to coalesce […] and the majority of Web pages are decorated with some form of semantic hypertext.

By 2012, 80% of public Web sites will use some level of semantic hypertext to create SW documents […] 15% of public Web sites will use more extensive Semantic Web-based ontologies to create semantic databases

(note: "semantic hypertext" refers to pages "prepared" for integration)

"Finding and Exploiting Value in Semantic Web Technologies on the Web", Gartner Research Report, May 2007

The "corporate" landscape is moving

- Major companies offer (or will offer) Semantic Web tools or systems using Semantic Web: Adobe, Oracle, IBM, HP, Software AG, GE, Northrop Gruman, Altova, Microsoft, Dow Jones, …
- Others are using it (or consider using it) as part of their own operations: Novartis, Pfizer, Telefónica, …
- Some of the names of active participants in W3C SW related groups: ILOG, HP, Agfa, SRI International, Fair Isaac Corp., Oracle, Boeing, IBM, Chevron, Siemens, Nokia, Pfizer, Sun, Eli Lilly, …

Lots of Tools (not an exhaustive list!)

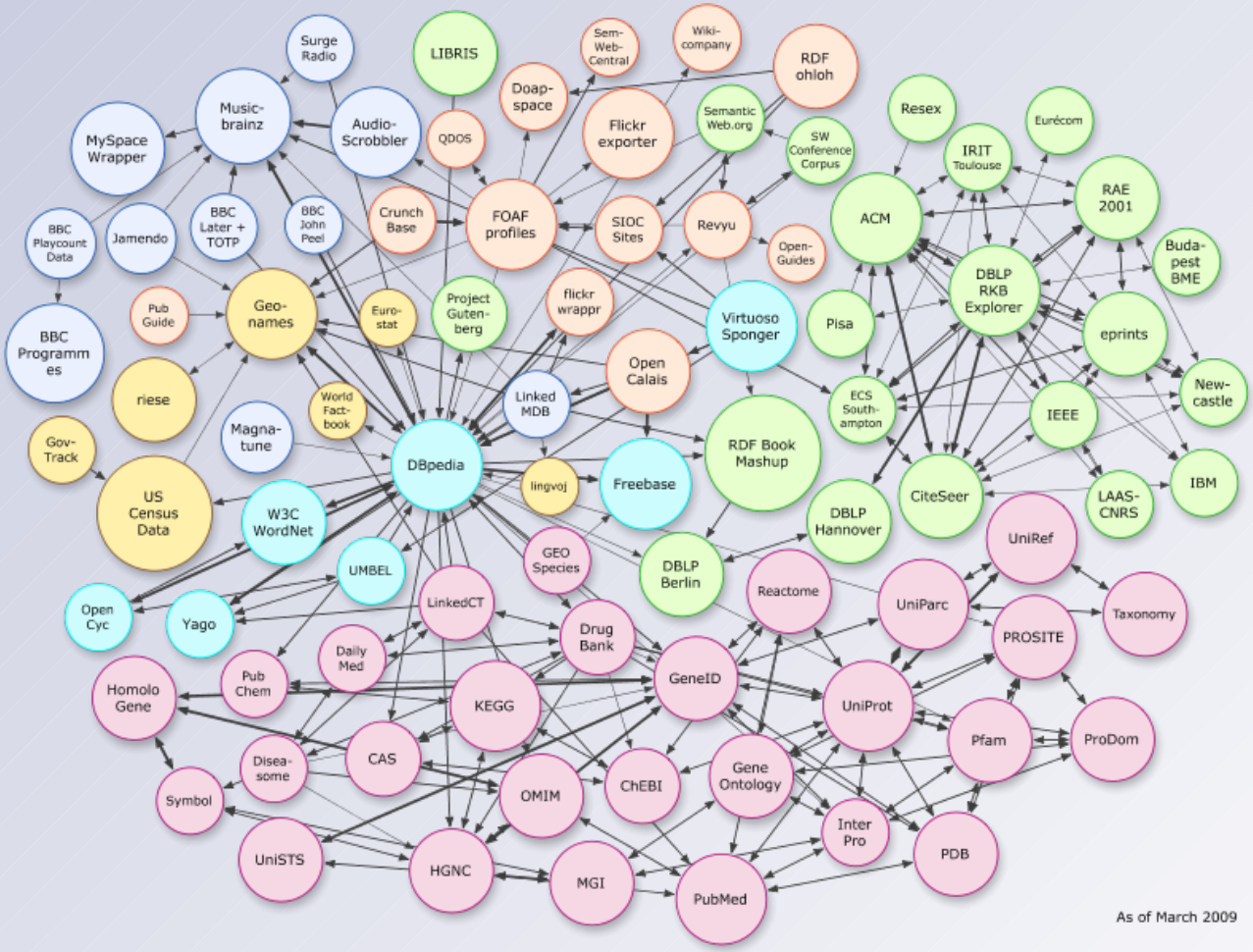
- Categories:
  - Triple Stores
  - Inference engines
  - Converters
  - Search engines
  - Middleware
  - CMS
  - Semantic Web browsers
  - Development environments
  - Semantic Wikis
  - …
- Some names:
  - Jena, AllegroGraph, Mulgara, Sesame, flickurl, …
  - TopBraid Suite, Virtuoso environment, Falcon, Drupal 7, Redland, Pellet, …
  - Disco, Oracle 11g, RacerPro, IODT, Ontobroker, OWLIM, Tallis Platform, …
  - RDF Gateway, RDFLib, Open Anzo, DartGrid, Zitgist, Ontotext, Protégé, …
  - Thetus publisher, SemanticWorks, SWI-Prolog, RDFStore…
  - …

Some deployment communities

- Major communities pick the technology up: digital libraries, defence, eGovernment, energy sector, financial services, health care, oil and gas industry, life sciences …
  - Health care and life science sector is now very active
    - also at W3C, in the form of an Interest Group

Application specific portions of the cloud

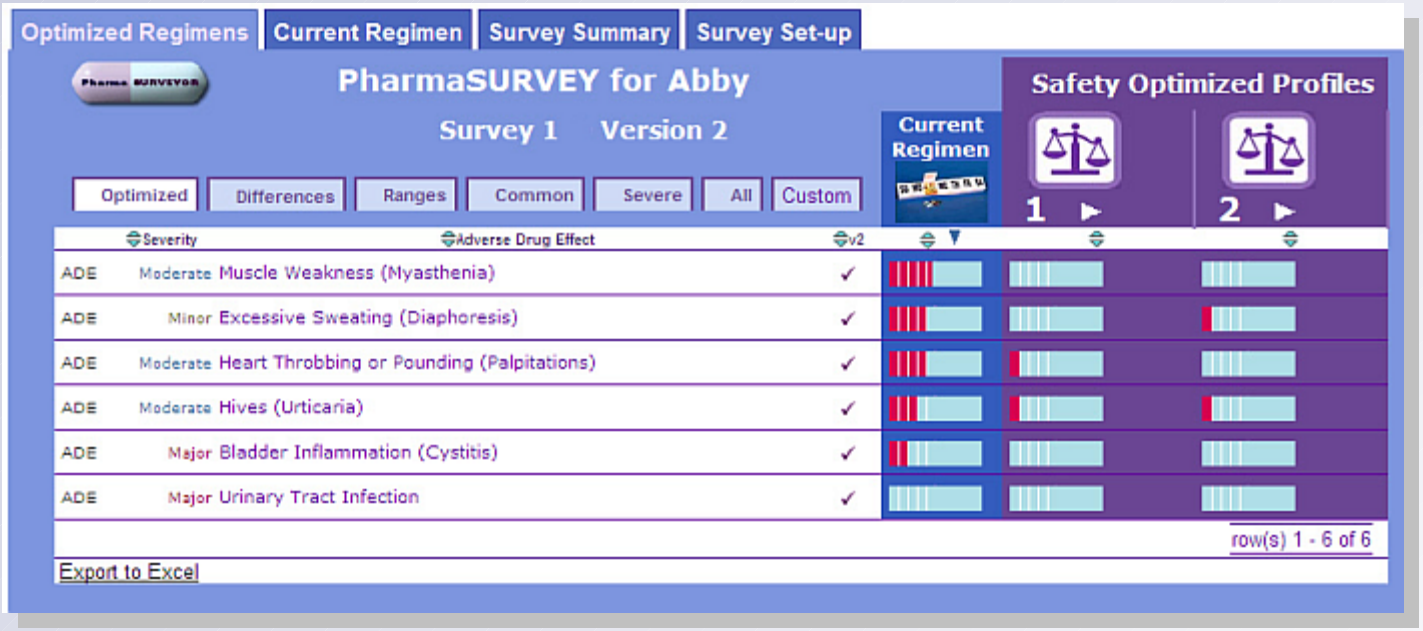
- Eg, “bio” related datasets
  - done, partially, by the “Linking Open Drug Data” task force of the HCLS IG at W3C



W3C Semantic Web

Help in choosing the right drug regimen

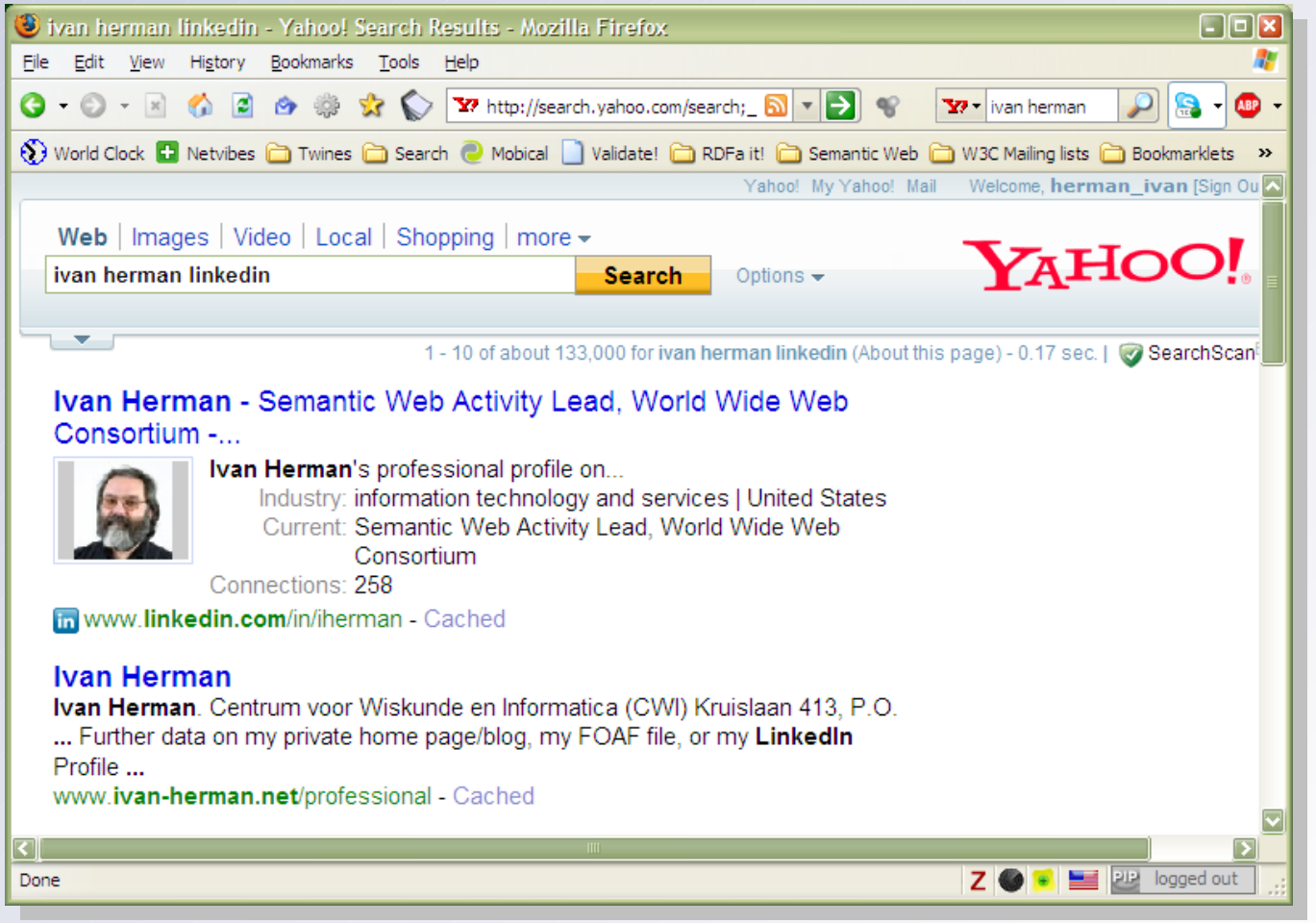
- Help in finding the best drug regimen for a specific case, per patient
- Integrate data from various sources (patients, physicians, Pharma, researchers, ontologies, etc)
- Data (eg, regulation, drugs) change often, but the tool is much more resistant against change



Courtesy of Erick Von Schweber, PharmaSURVEYOR Inc., (SWEO Use Case)

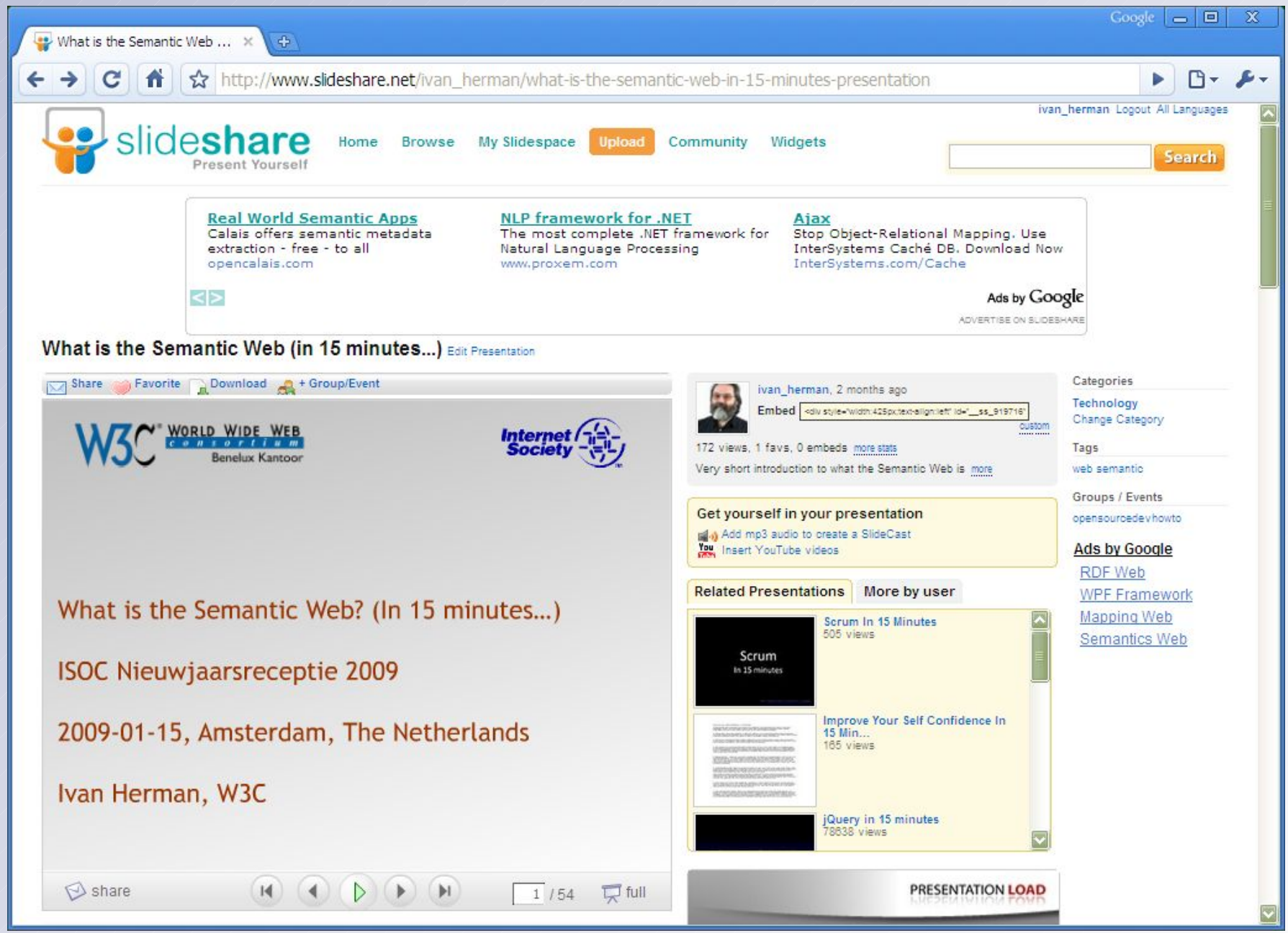
Yahoo's SearchMonkey

- Search based results may be customized via small applications
- Metadata embedded in pages are reused
- Publishers can export extra data via other formats



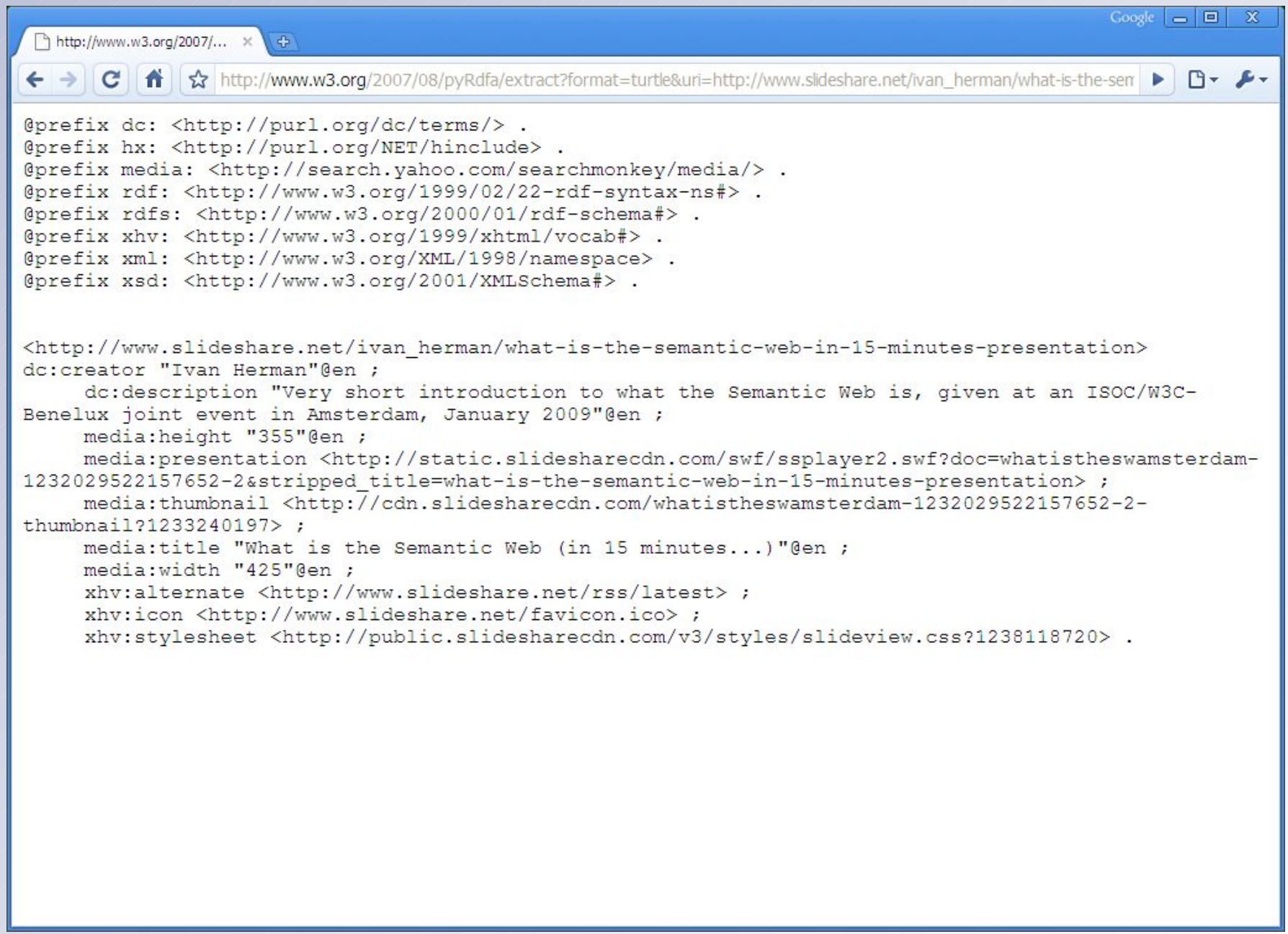
Courtesy of Peter Mika, Yahoo! Research, (SWEO Case Study)

Information in Web Pages: SlideShare



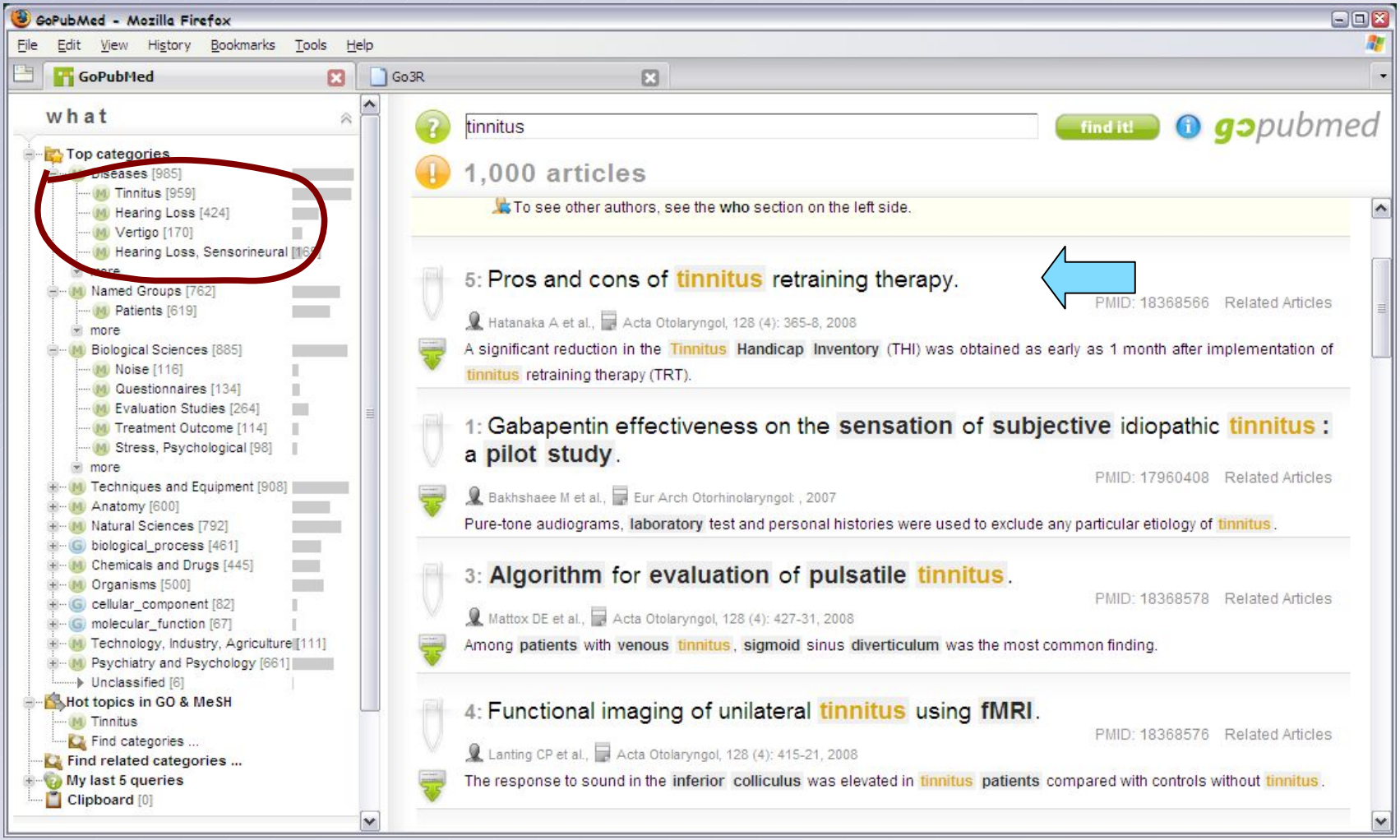
W3C Semantic Web

Information in Web Pages: SlideShare



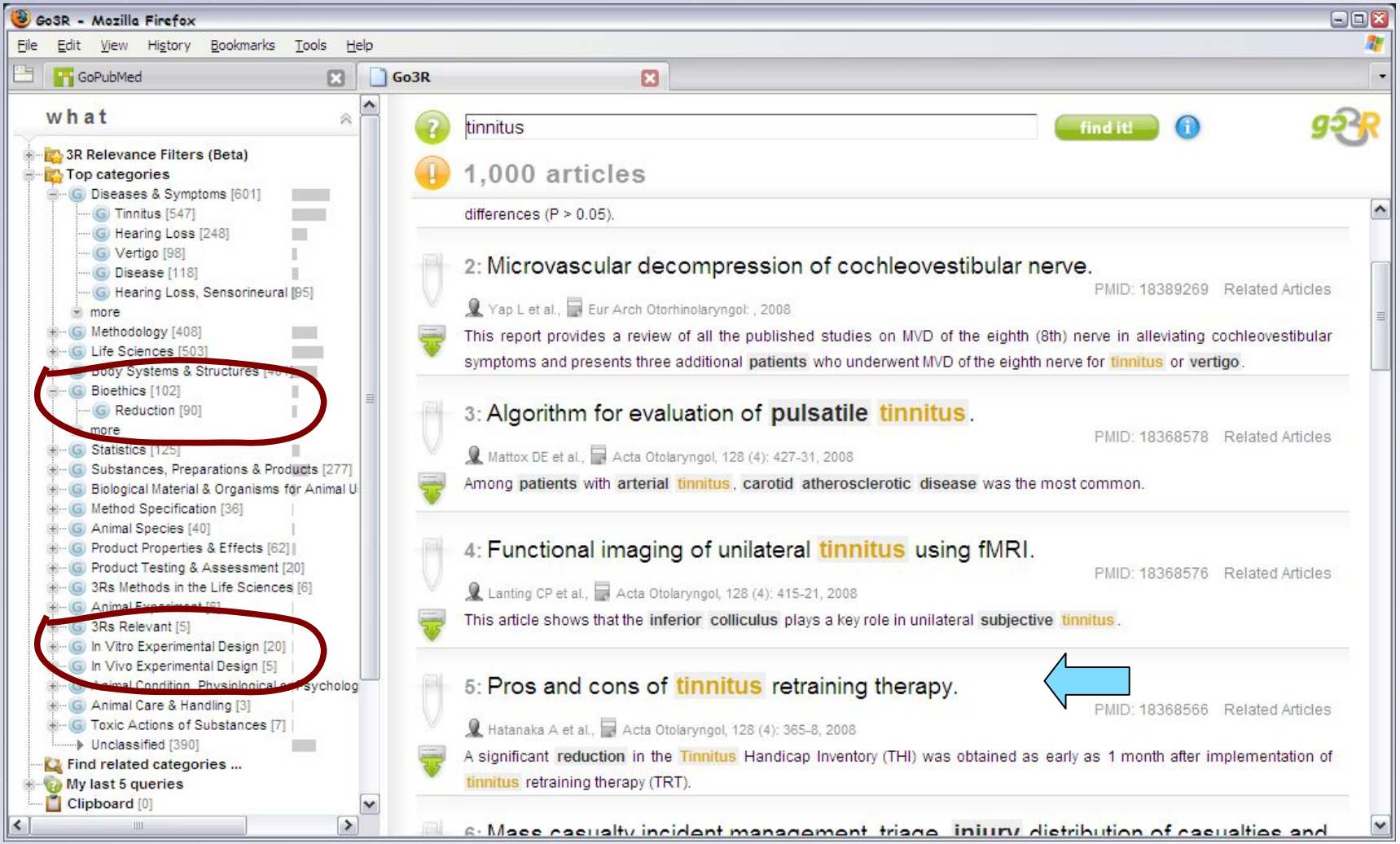
Improved Search (GoPubMed)

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



Improved Search (Go3R)

- Same dataset, different ontology
  - (ontology is on non-animal experimentation)



New type of Web 2.0 applications

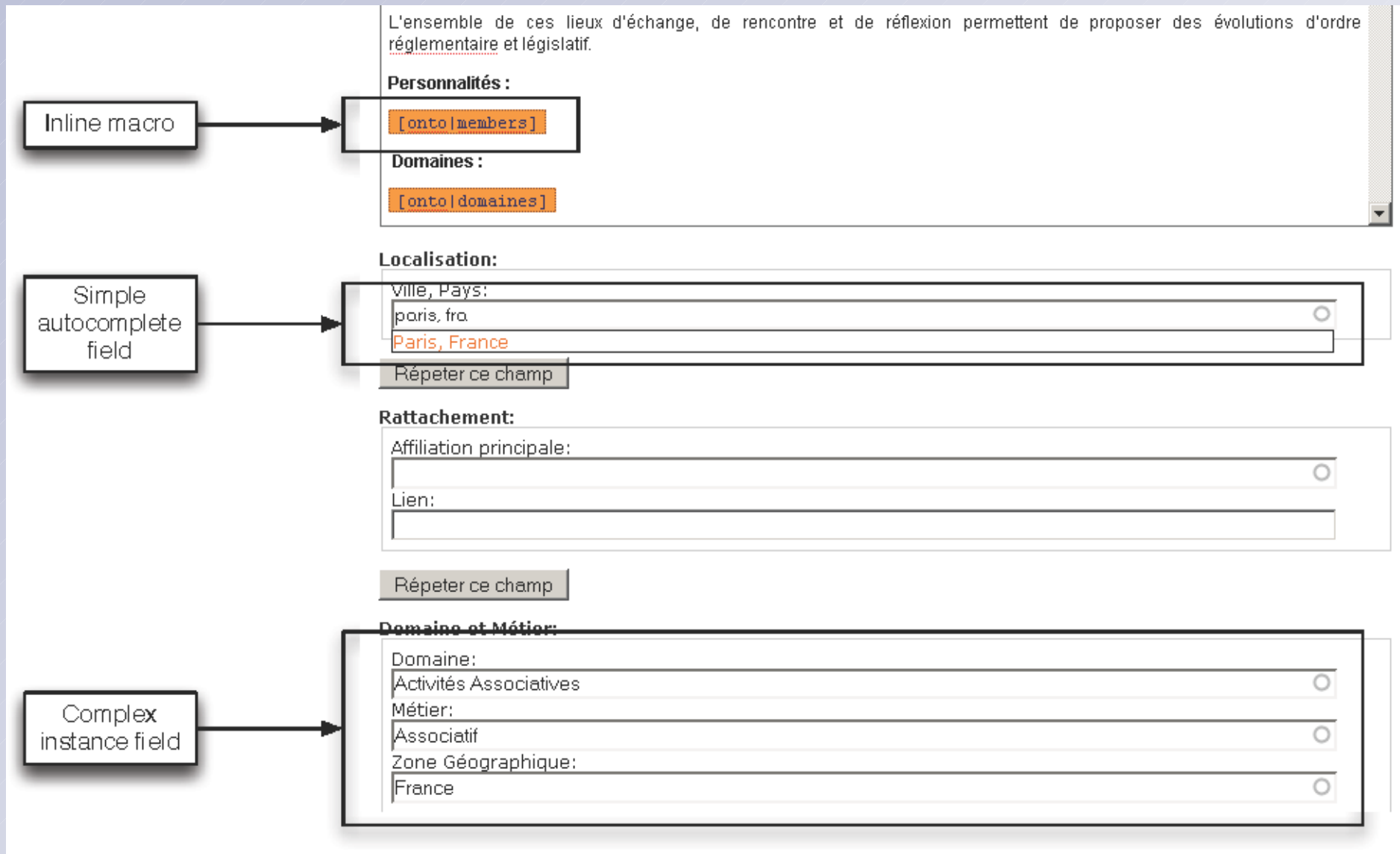
- New Web 2.0 applications come every day
- Some begin to look at Semantic Web as possible technology to improve their operation
  - more structured tagging, making use of external services
  - providing extra information to users
  - etc.
- Some examples: Twine, Revyu, Faviki, …

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
  - Semantic Web based unifying format
  - simple, public vocabularies
  - internal data is combined with linked open data like Geonames
  - Semantic Web queries are is used for internally
- Details are hidden from end users (via plugins, extra layers, etc)

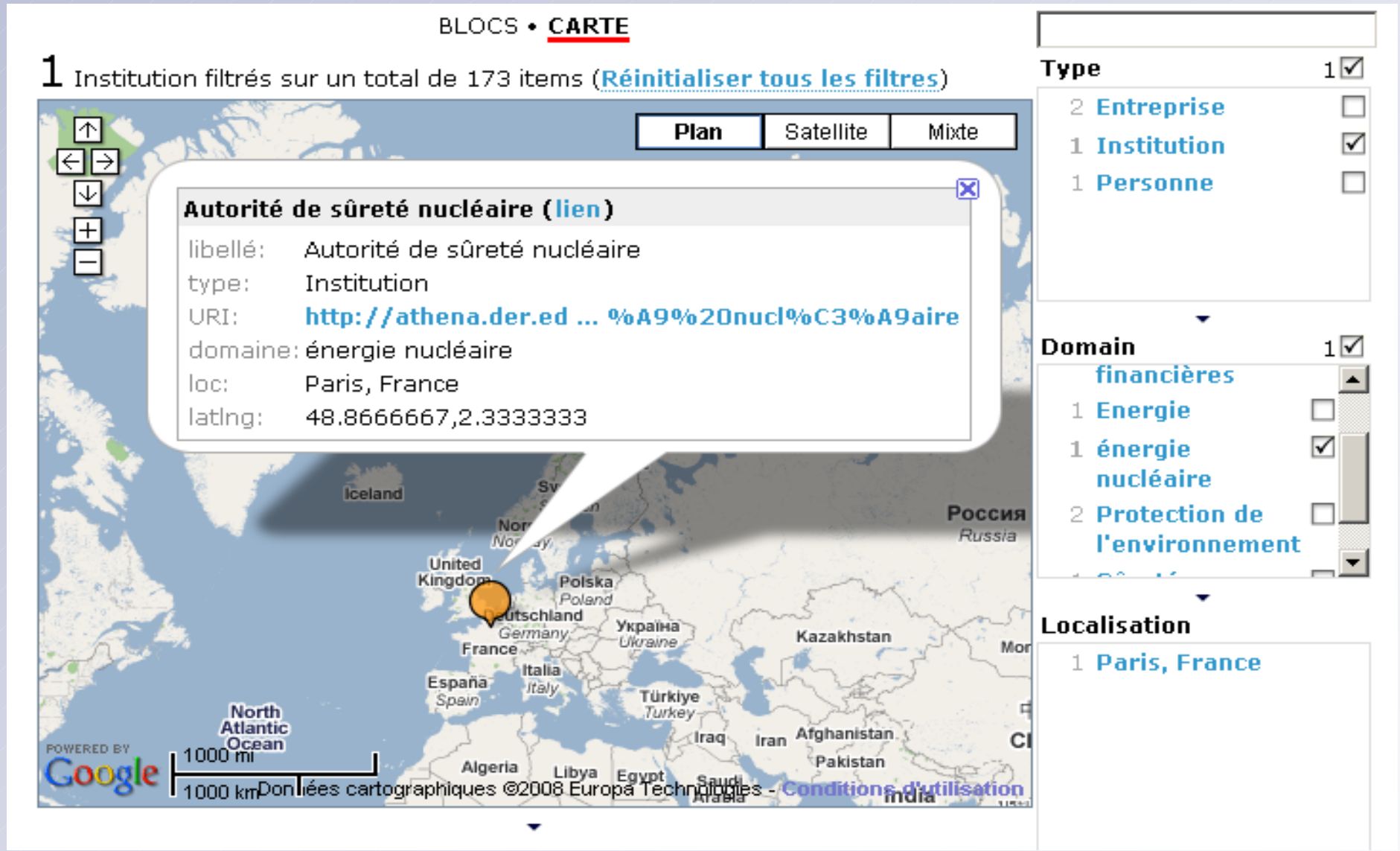
Courtesy of A. Passant, EDF R&D and LaLIC, Université Paris-Sorbonne, (SWEO Case Study)

Integration of “social” software data



Courtesy of A. Passant, EDF R&D and LaLIC, Université Paris-Sorbonne, (SWEO Case Study)

Integration of “social” software data



Courtesy of A. Passant, EDF R&D and LaLIC, Université Paris-Sorbonne, (SWEO Case Study)

Conclusions…

- 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
  - …

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

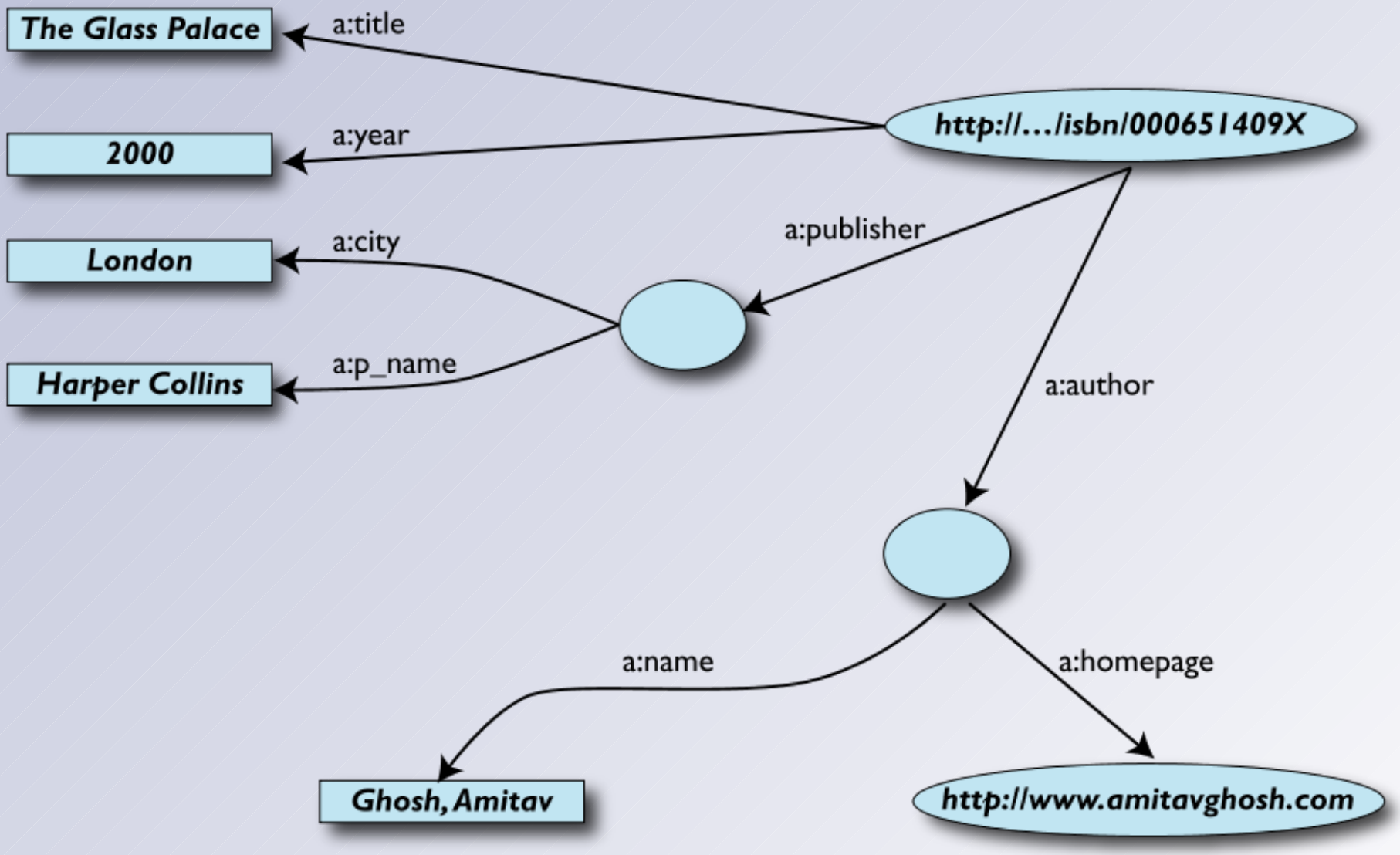
A *simplified* bookstore data (dataset "A")

ID	Author	Title	Publisher	Year
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

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

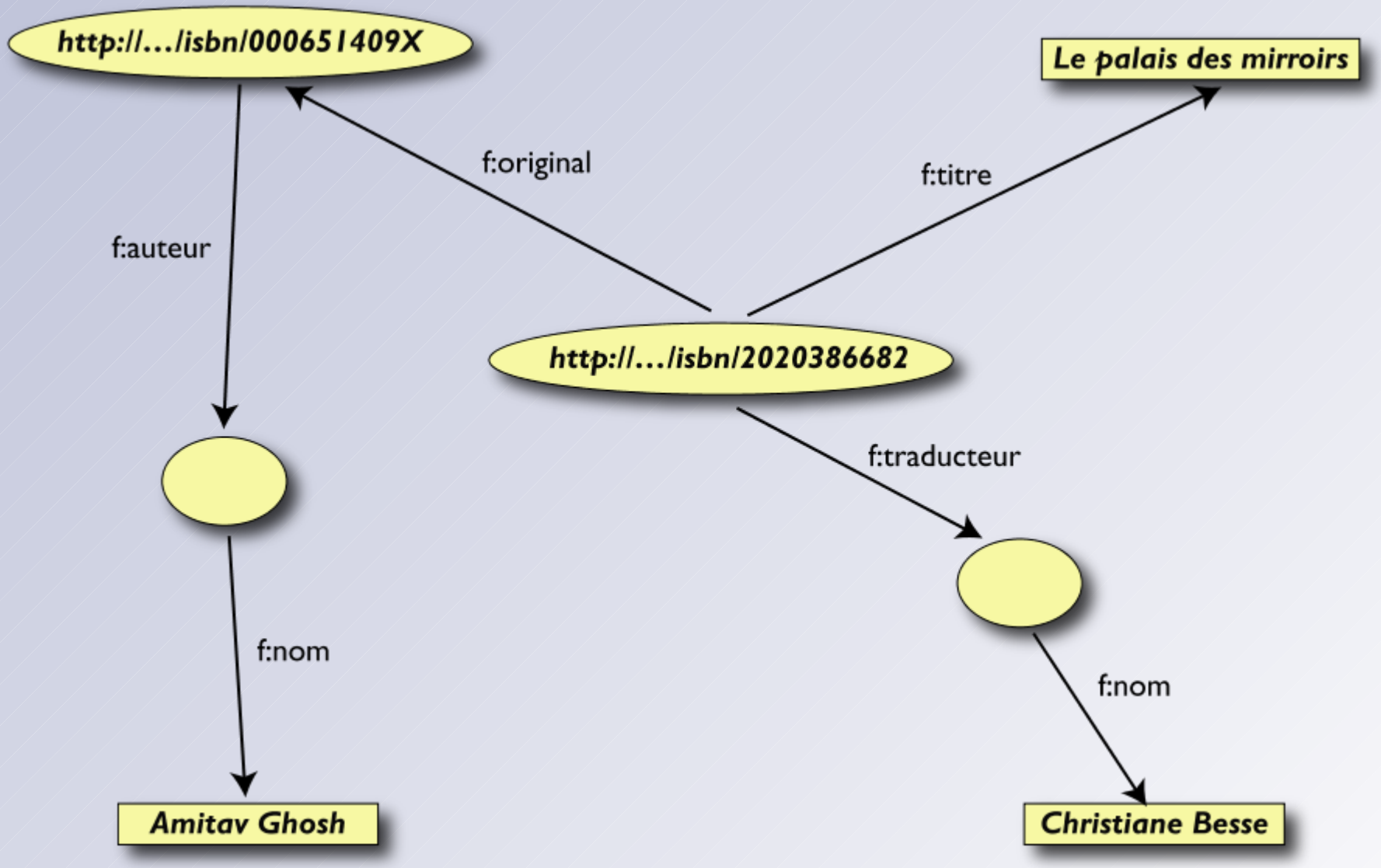
Another *bookstore data (dataset "F")*

	A	B	D	E
1	**ID**	**Titre**	**Traducteur**	**Original**
2	ISBN0 2020386682	Le Palais des miroirs	A13	ISBN-0-00-651409-X
3				

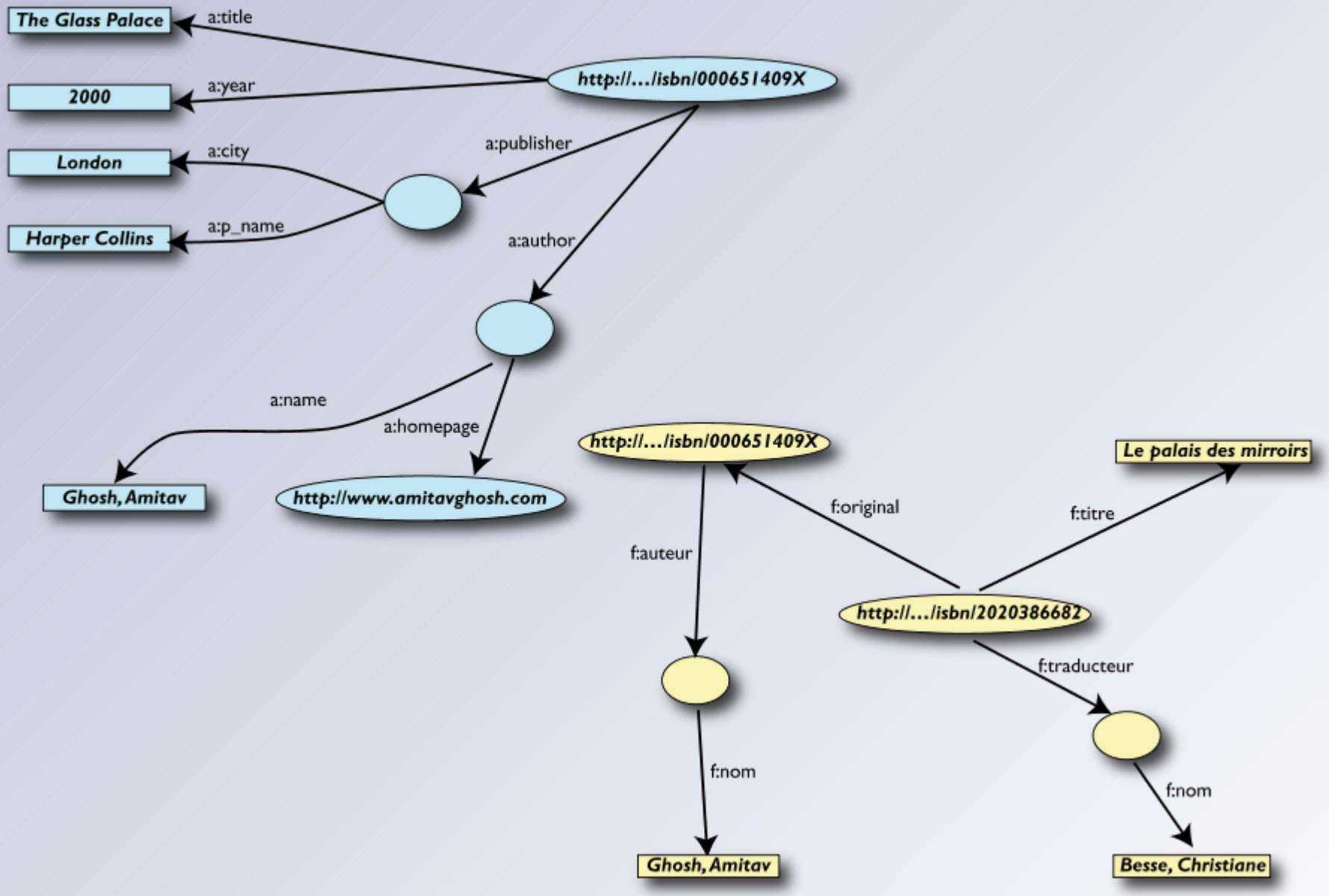
6	**ID**	**Auteur**
7	ISBN-0-00-651409-X	A12

11	**Nom**
12	Ghosh, Amitav
13	Besse, Christianne

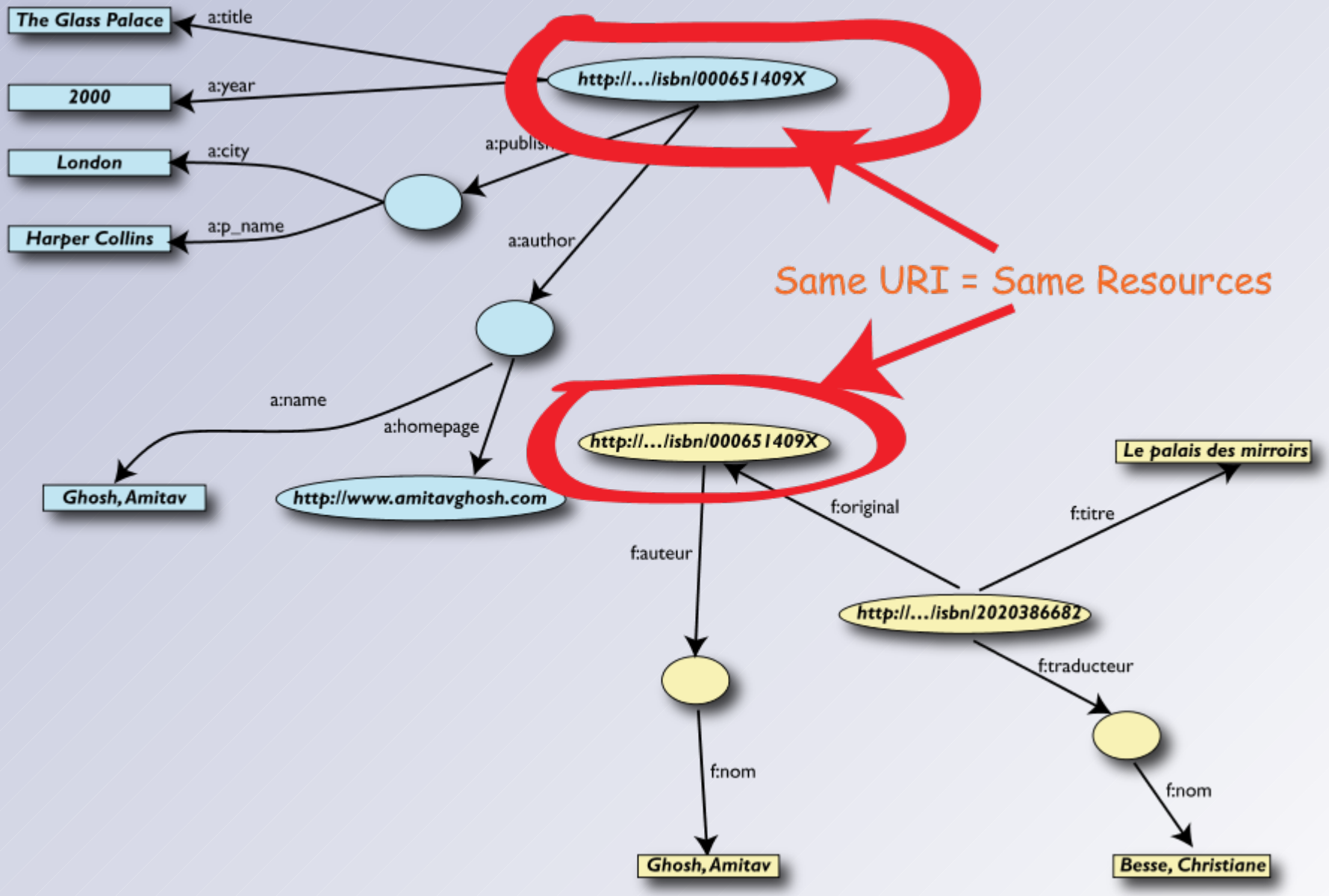
2nd: export your second set of data



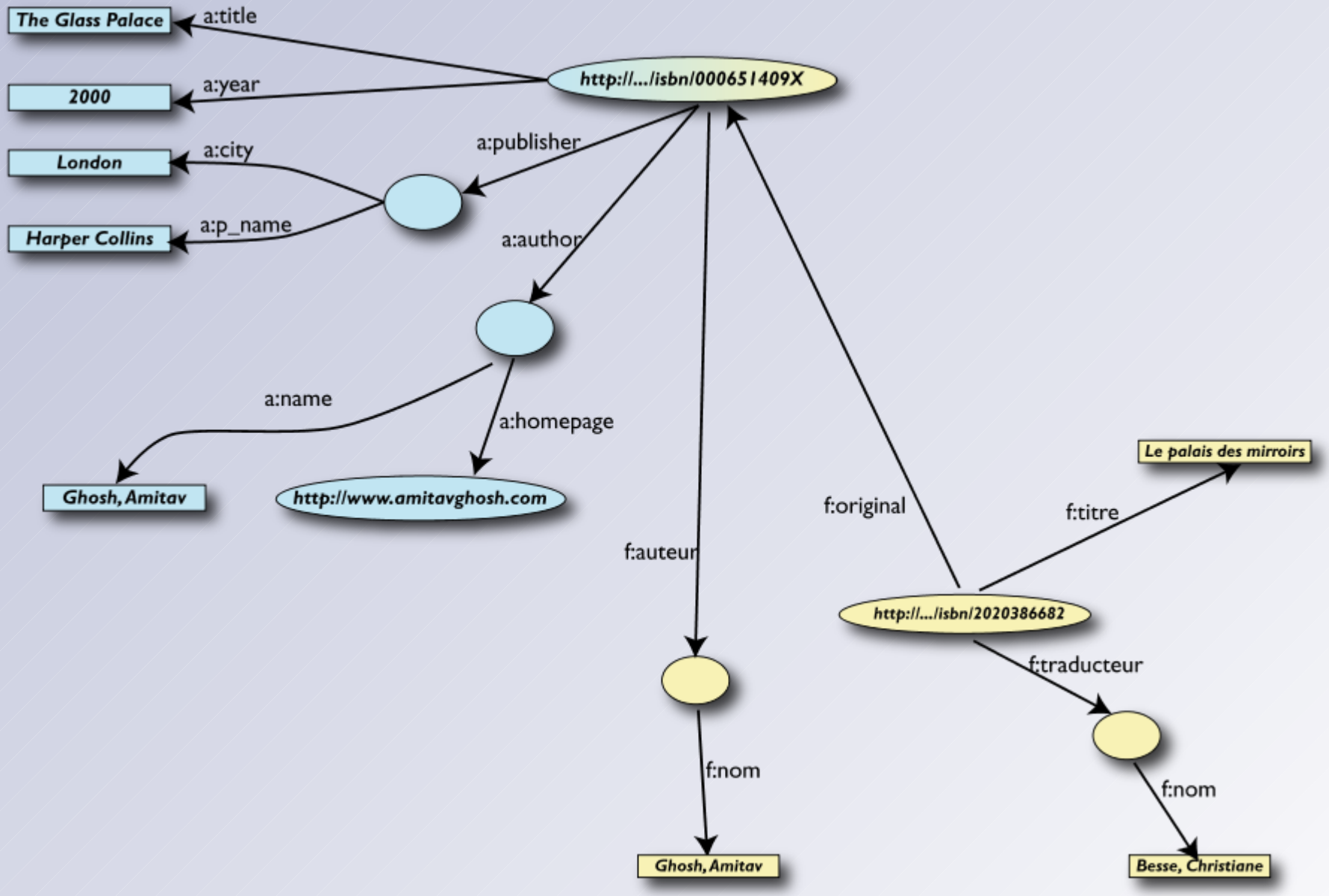
3rd: start merging your data



3rd: start merging your data (cont.)

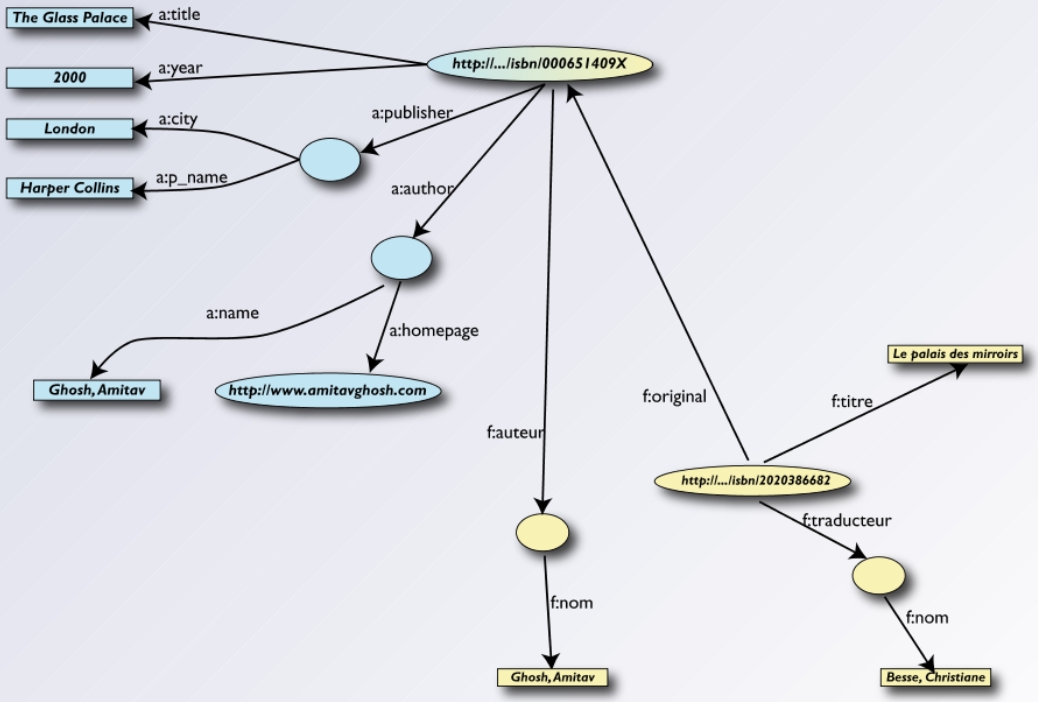


3rd: merge identical resources



Start making queries…

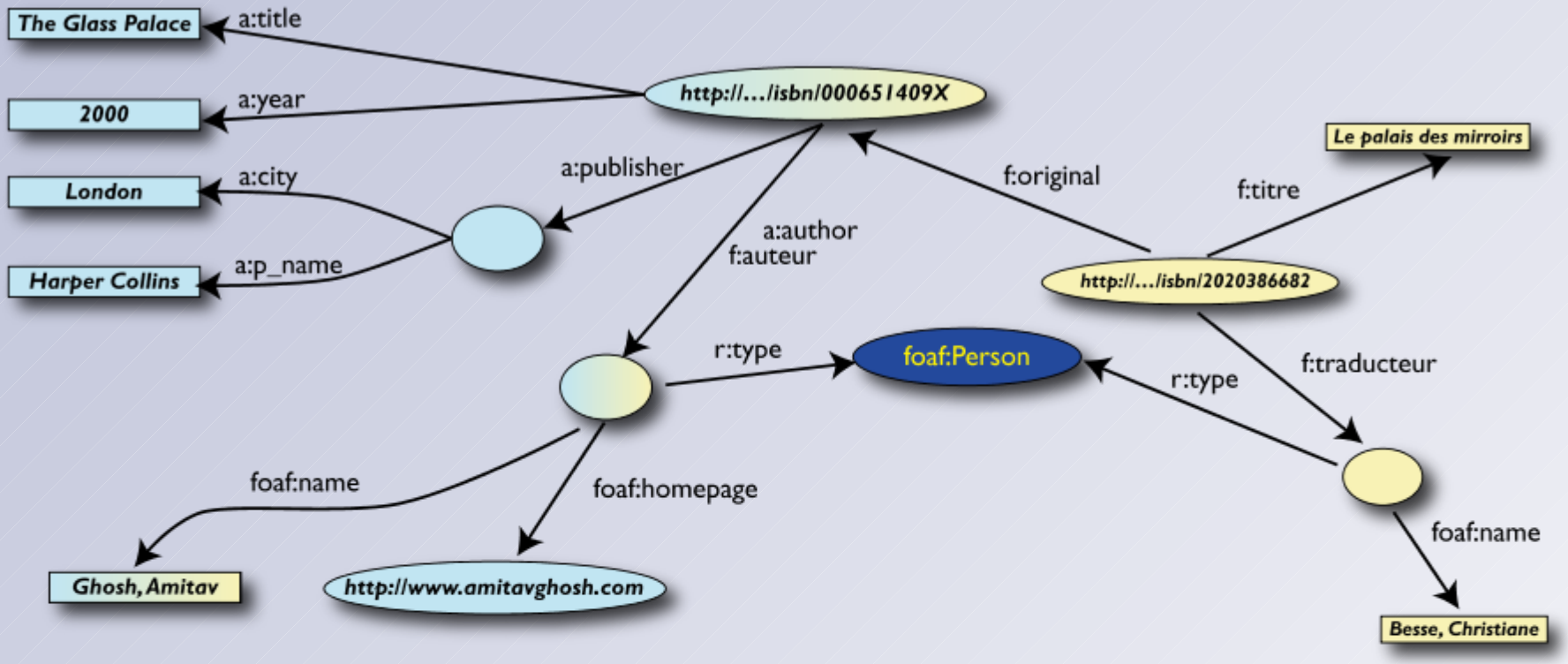
- User of data “F” can now ask queries like:
  - “give me the title of the original”
    - well, … « donnes-moi le titre de l’original »
- This information is not in the dataset “F”…
- …but can be retrieved by merging with dataset “A”!

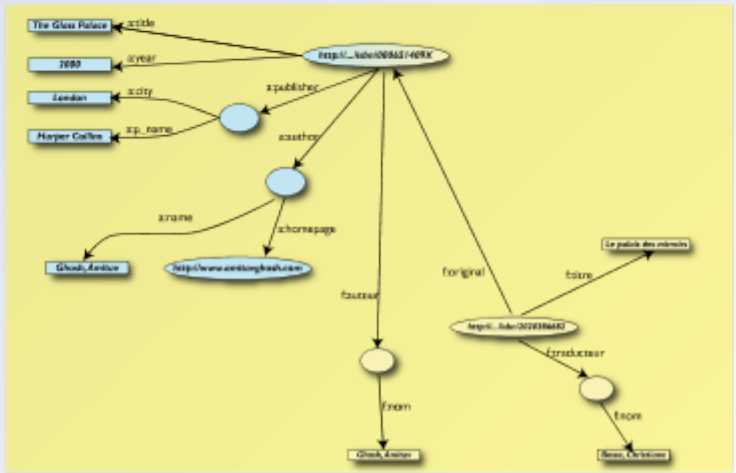


However, more can be achieved…

- 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

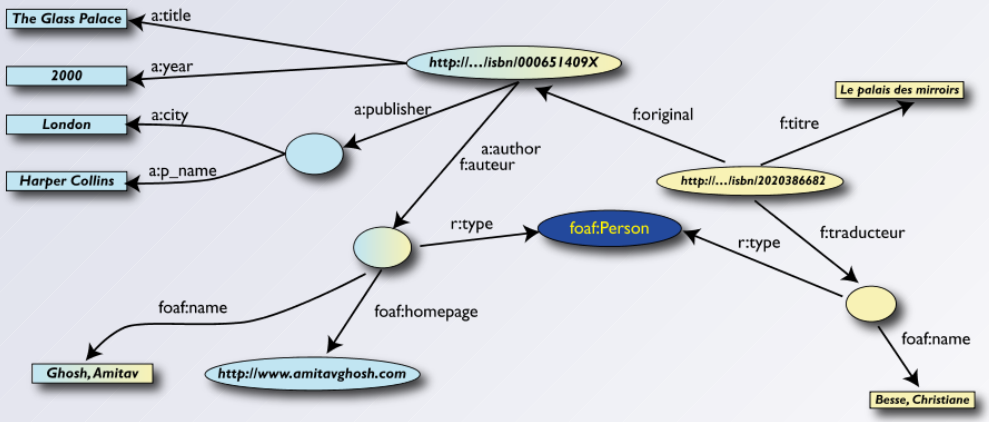
3rd revisited: use the extra knowledge





Start making richer queries!

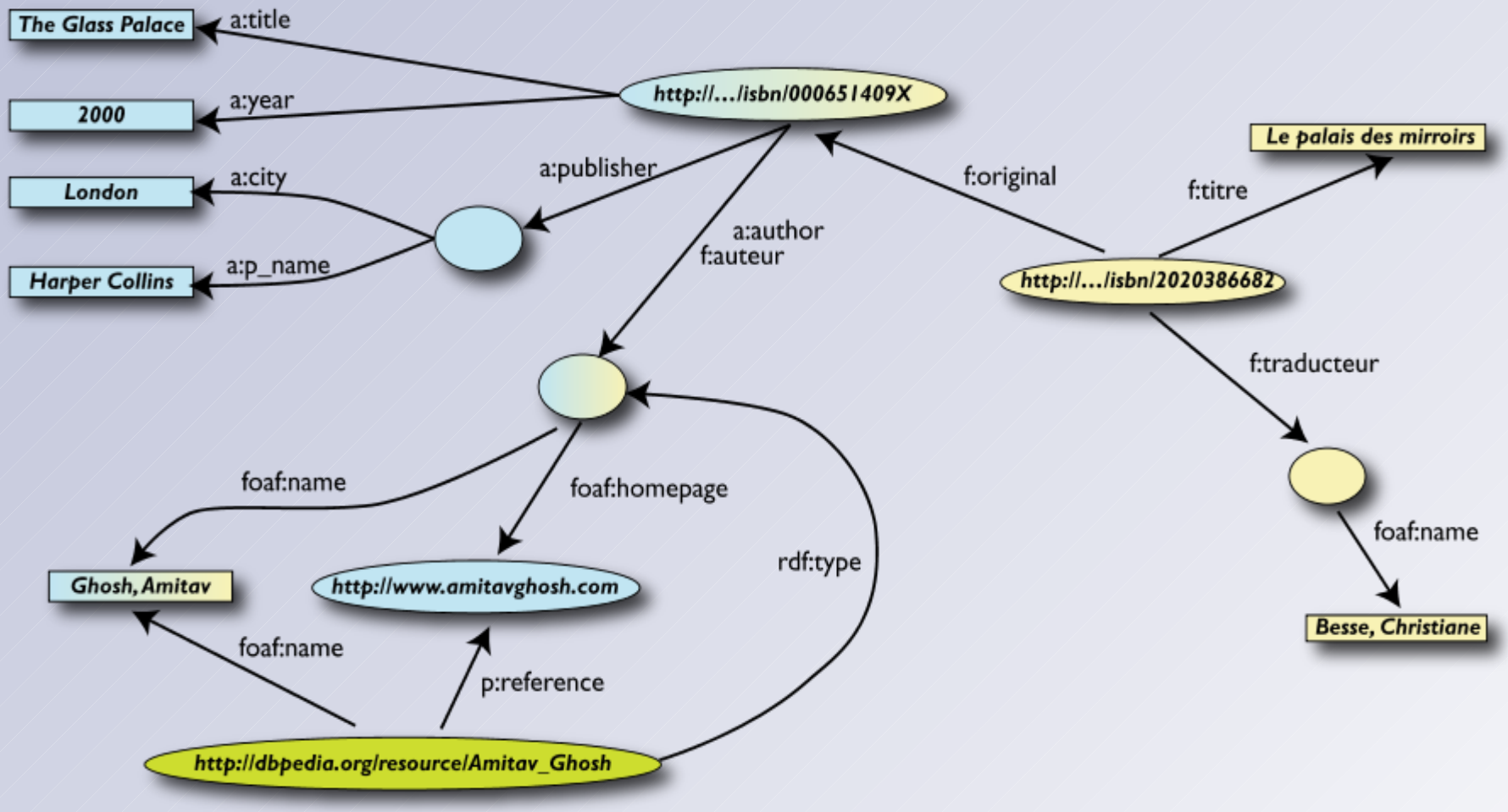
- User of dataset “F” can now query:
  - “donnes-moi la page d’accueil de l’auteur de l’originale”
  - well… “give me the home page of the original’s ‘auteur’”
- The information is not in datasets “F” or “A”…
- …but was made available by:
  - merging datasets “A” and datasets “F”
  - adding three simple extra statements as an extra “glue”



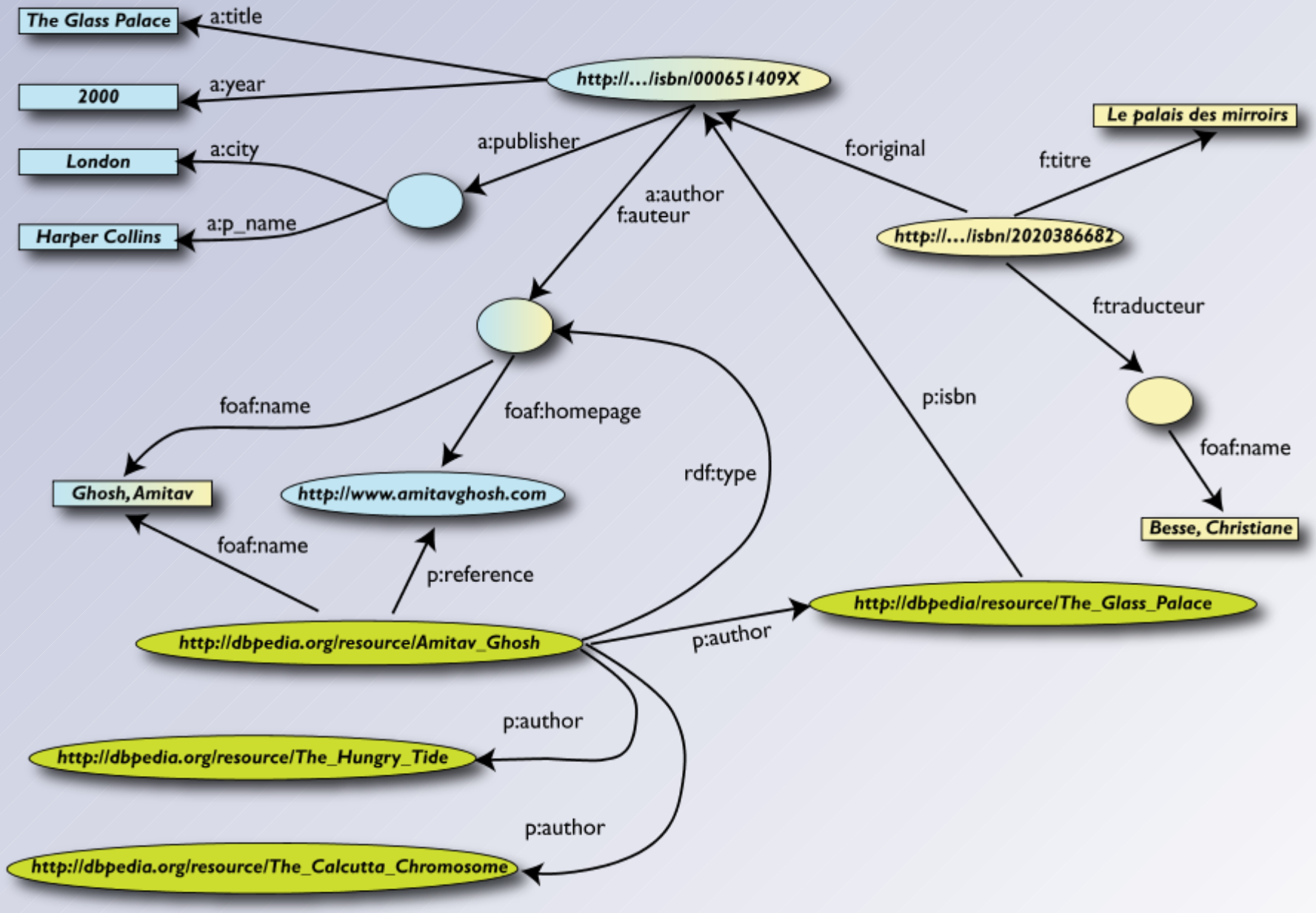
Combine with different datasets

- 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…

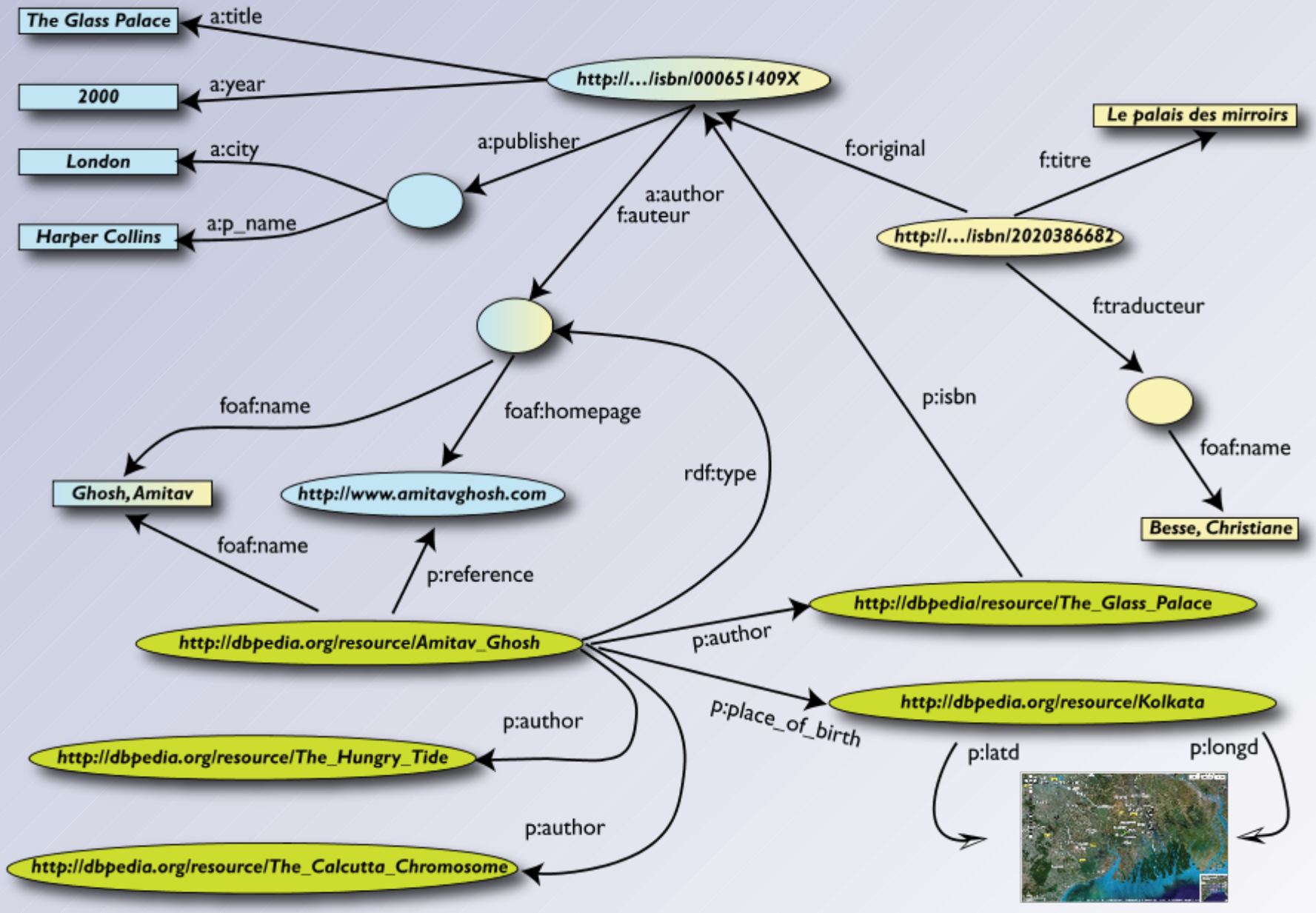
Merge with Wikipedia data



Merge with Wikipedia data



Merge with Wikipedia data



Is that surprising?

- It may look like it but, in fact, it should not be…
- What happened via automatic means is done every day by Web users!
- The difference: a bit of extra rigour so that machines could do this, too

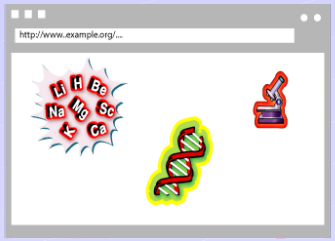
What did we do?

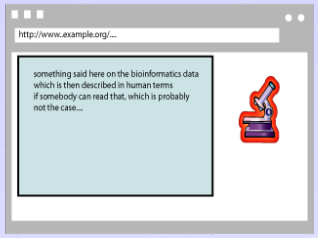
- We combined different datasets that
  - are somewhere on the web
  - are of different formats (mysql, excel sheet, XHTML, etc)
  - have different names for relations
- We could combine the data because some URI-s were identical (the ISBN-s in this case)
- We could add some simple additional information (the “glue”), possibly 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 *ontologies*, extra *rules*, 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

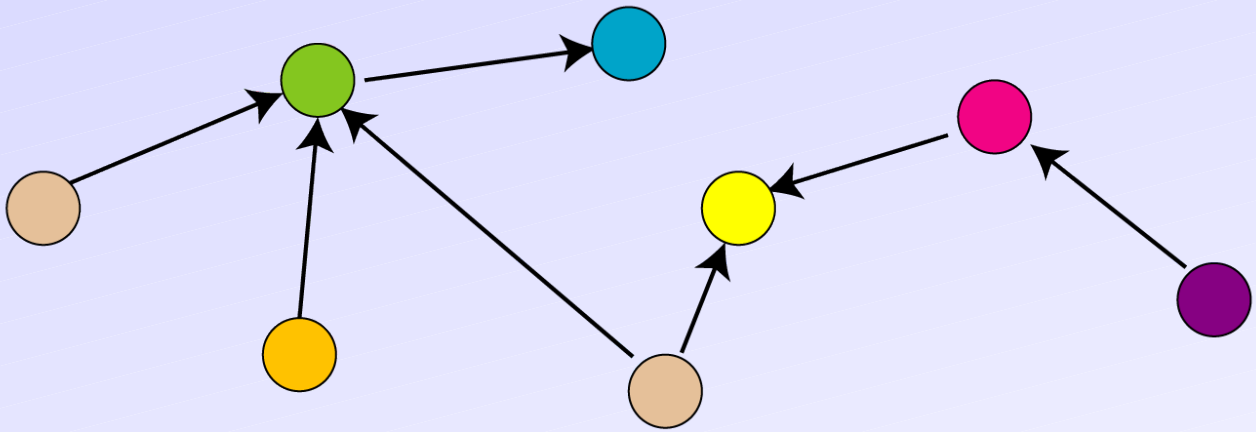
What did we do? (cont)





Applications

Query,
Manipulate,
etc.



Data represented in abstract format

Map,
Expose,
etc.

Data in various formats