What is the Semantic Web?

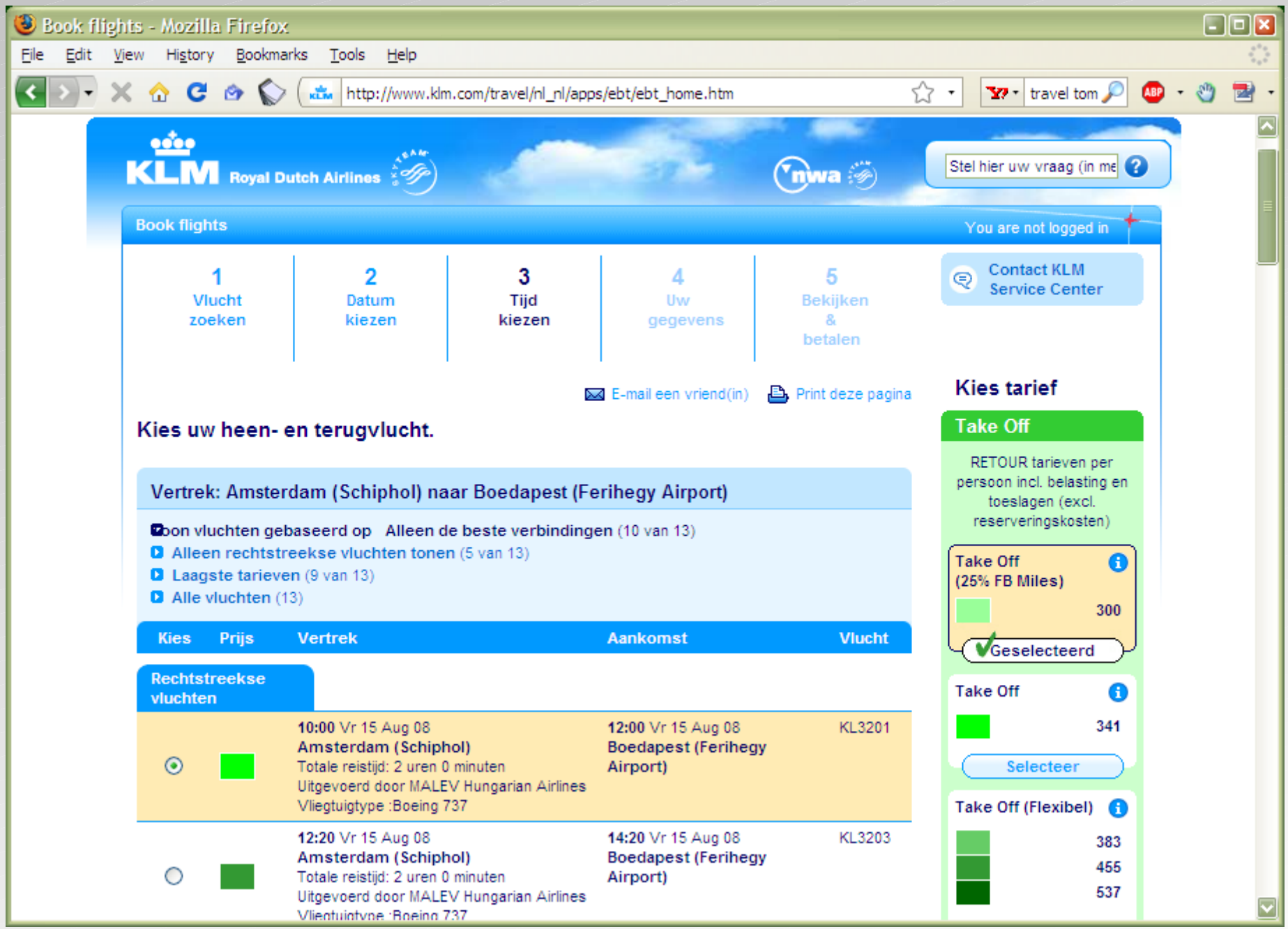
Ivan Herman, W3C

Last update: 2008-11-04

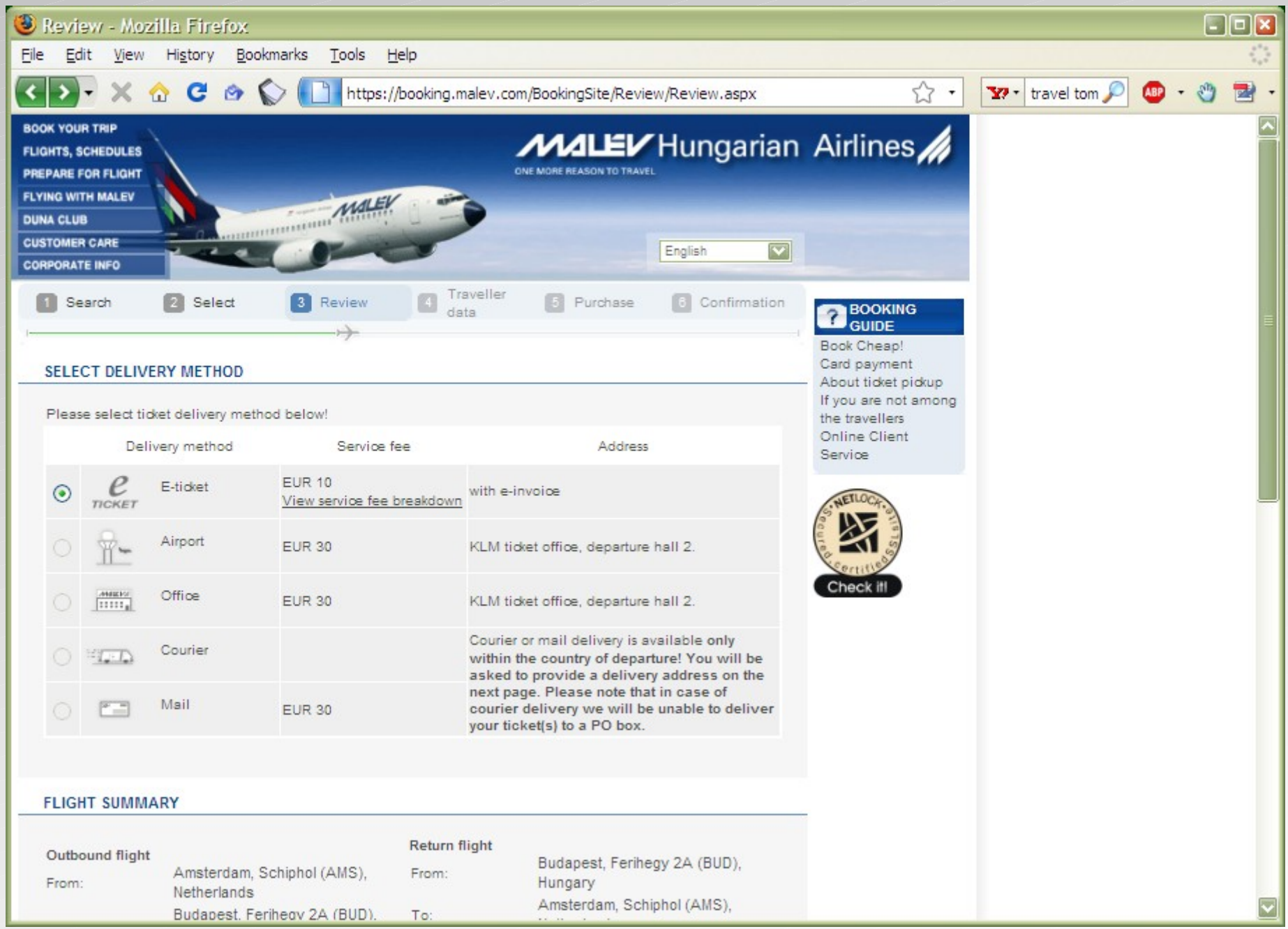
Let’s organize a trip to Budapest using the Web!

You try to find a proper flight with …

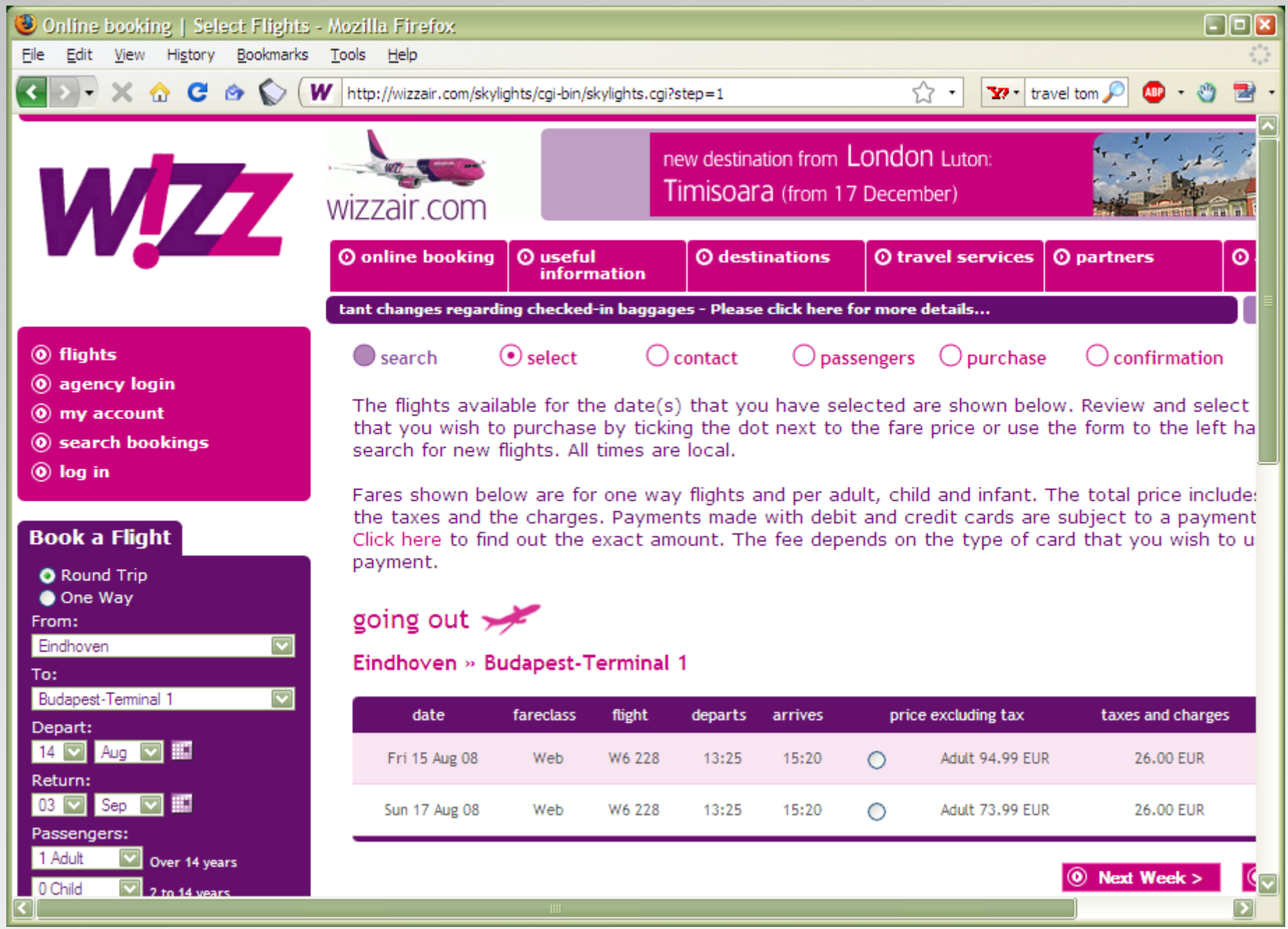
… a big, reputable airline, or …



… the airline of the target country, or …

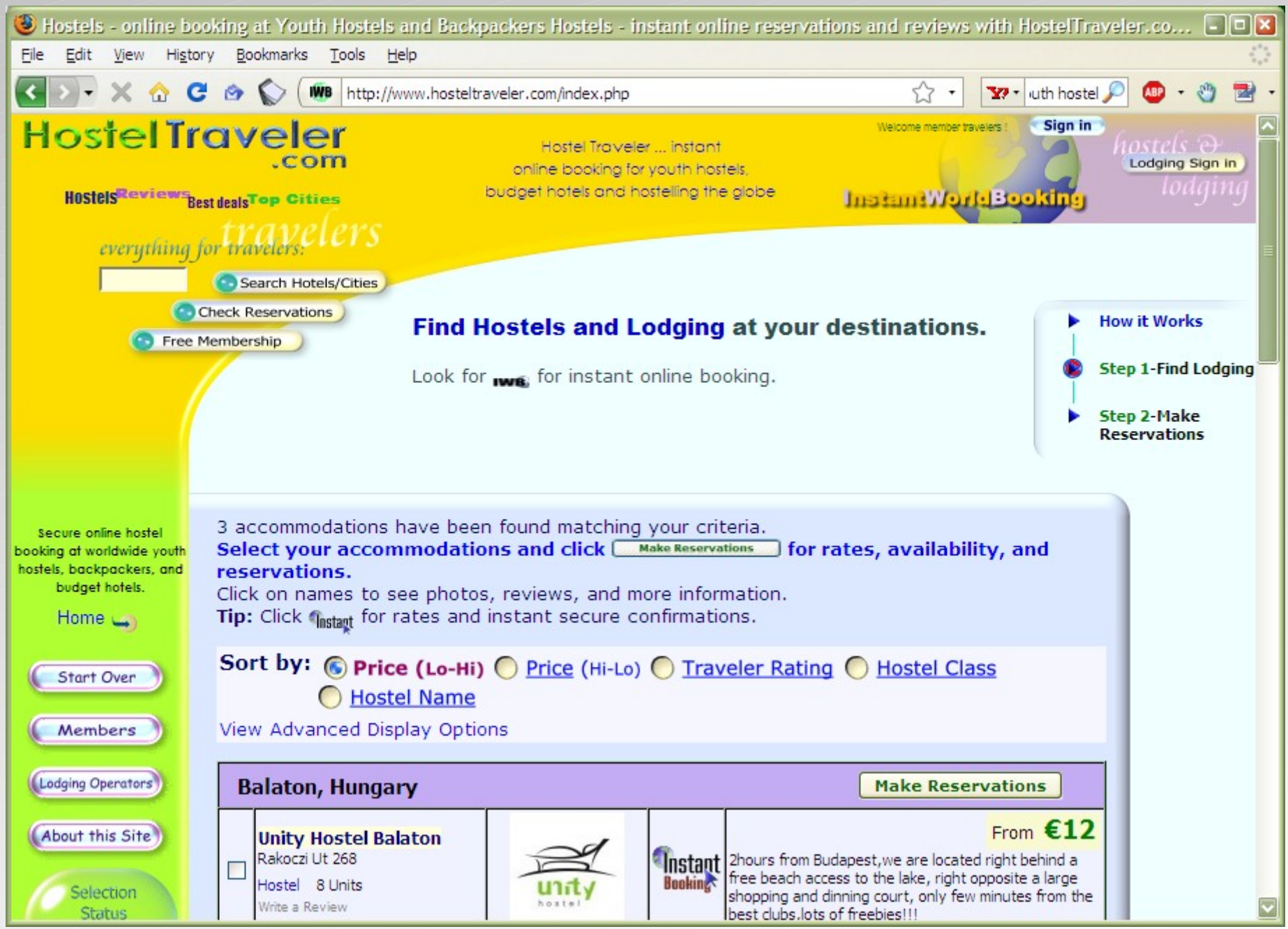


… or a low cost one

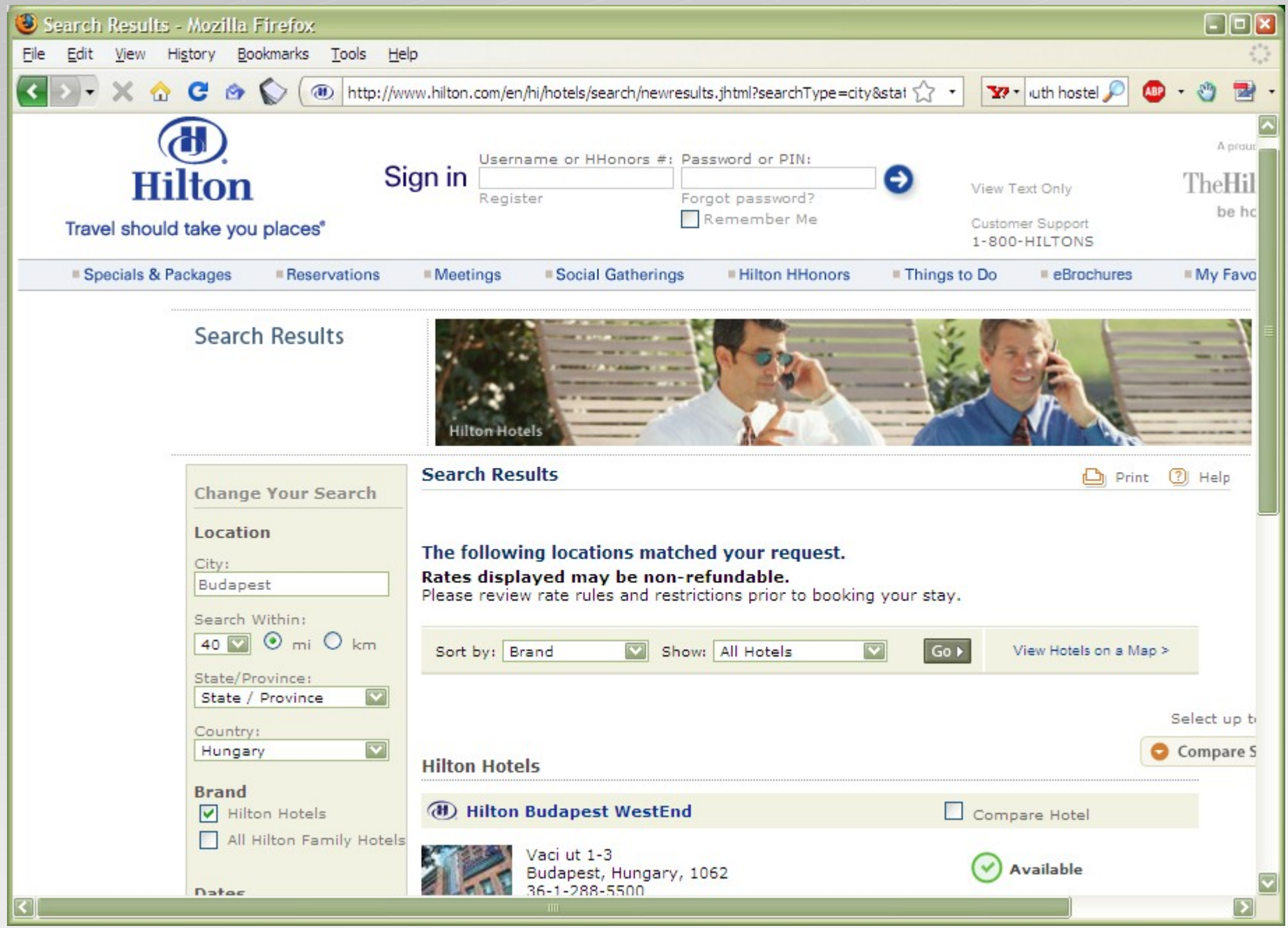


You have to find a hotel, so you look for…

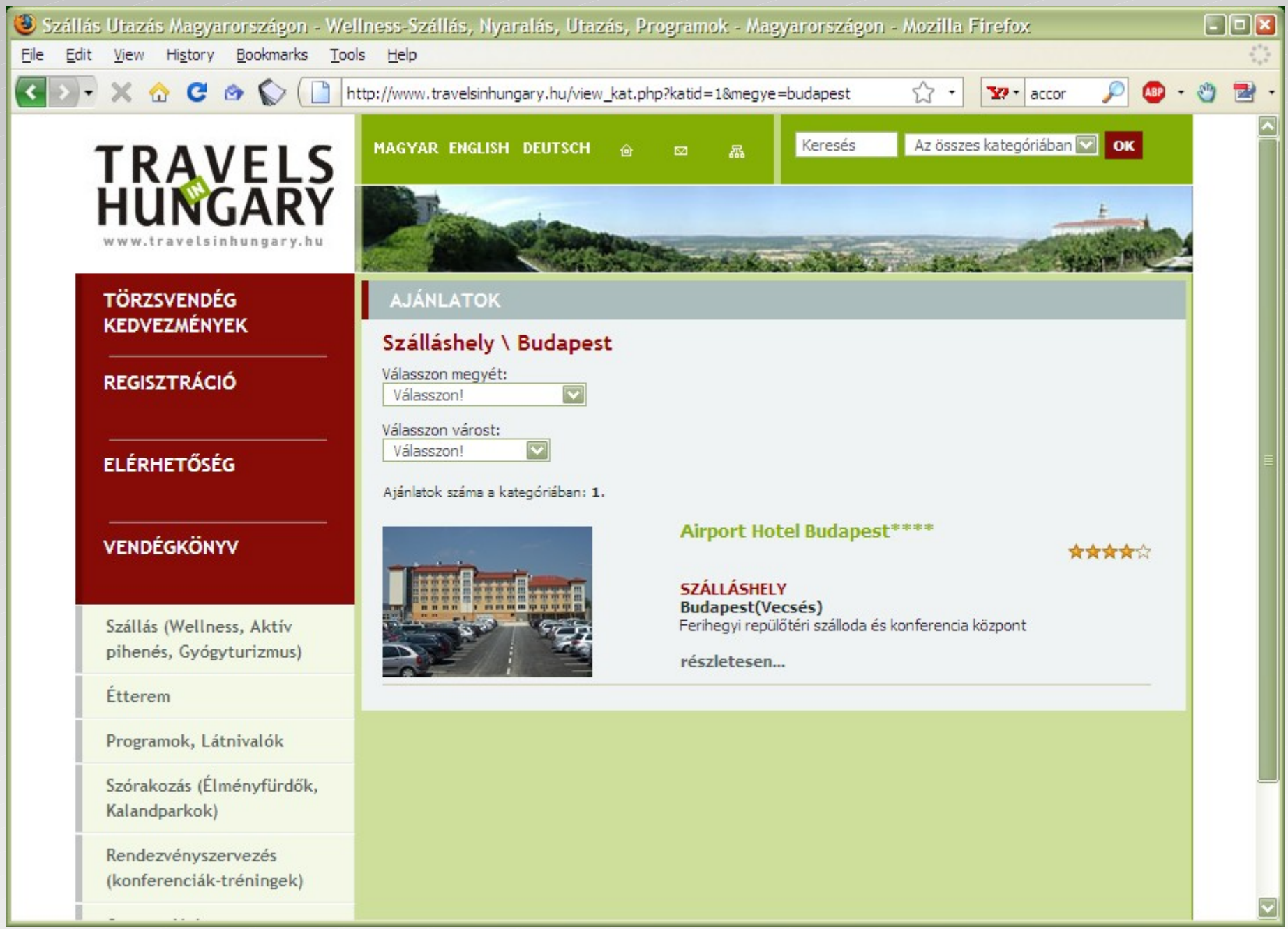
… a really cheap accommodation, or …



… or a really luxurious one, or …

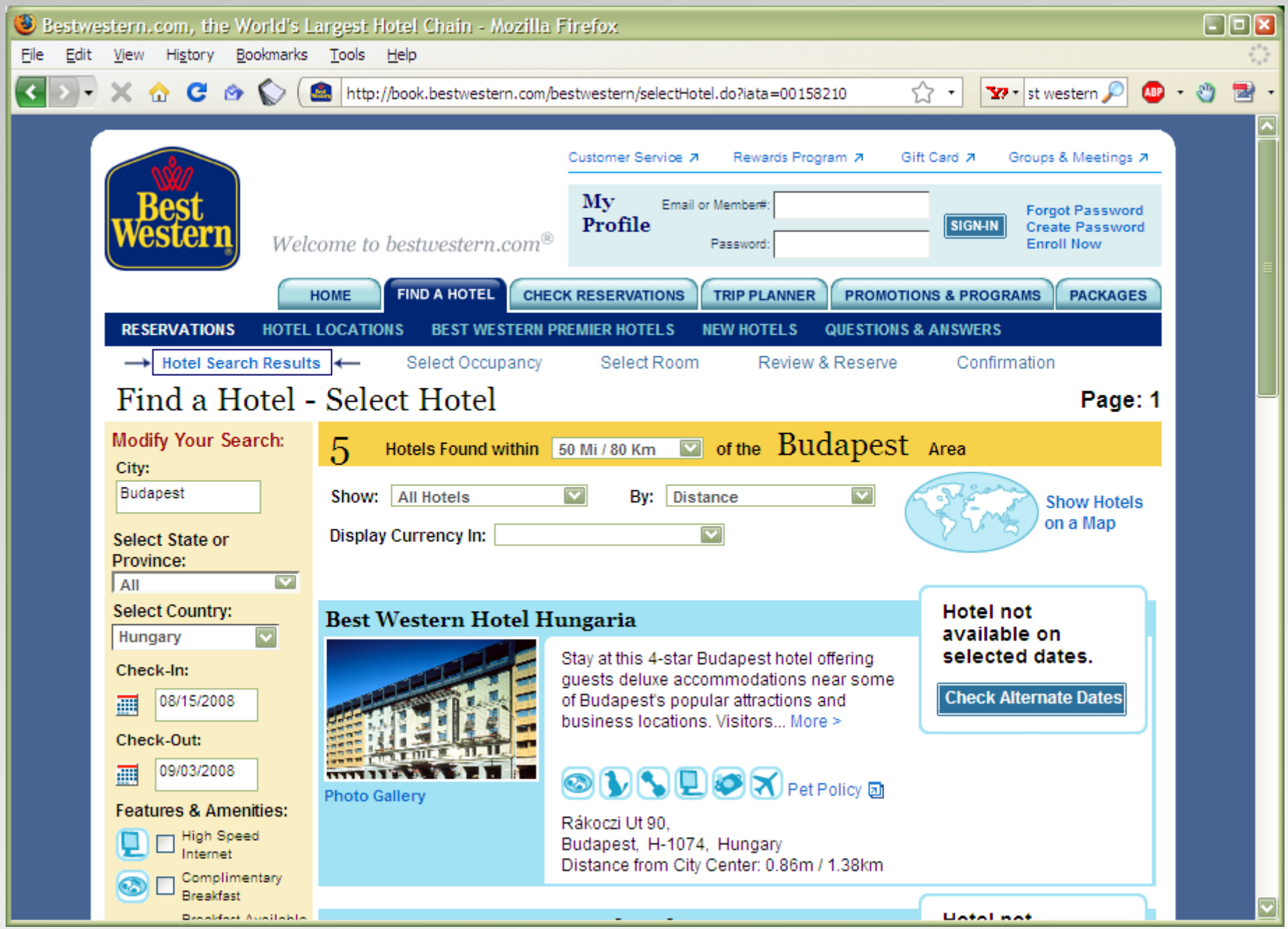


… and intermediate one …



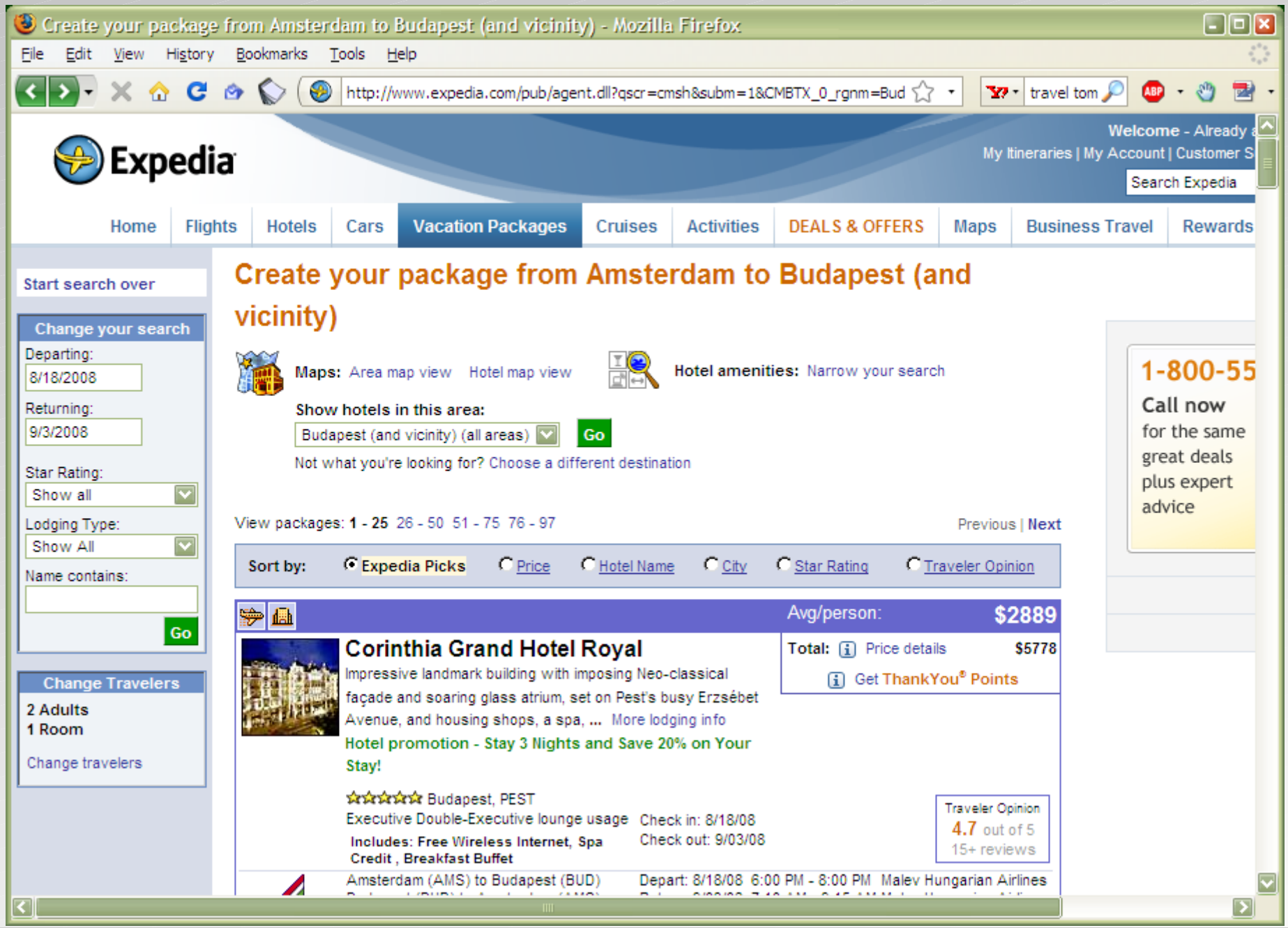
oops, that is no good, the page is in Hungarian that almost nobody understands, but…

… this one could work

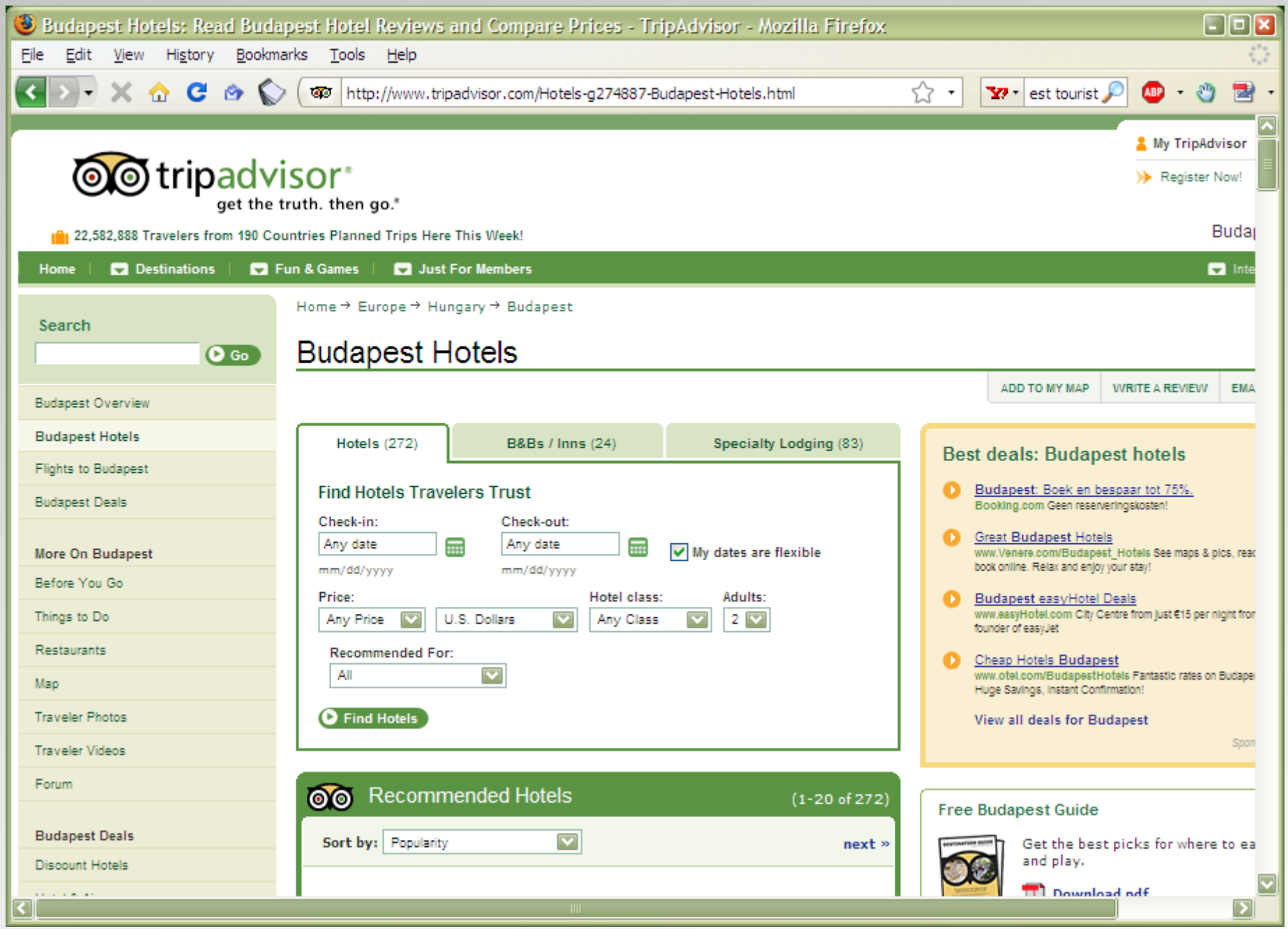


Of course, you could decide to trust a specialized site…

… like this one, or…

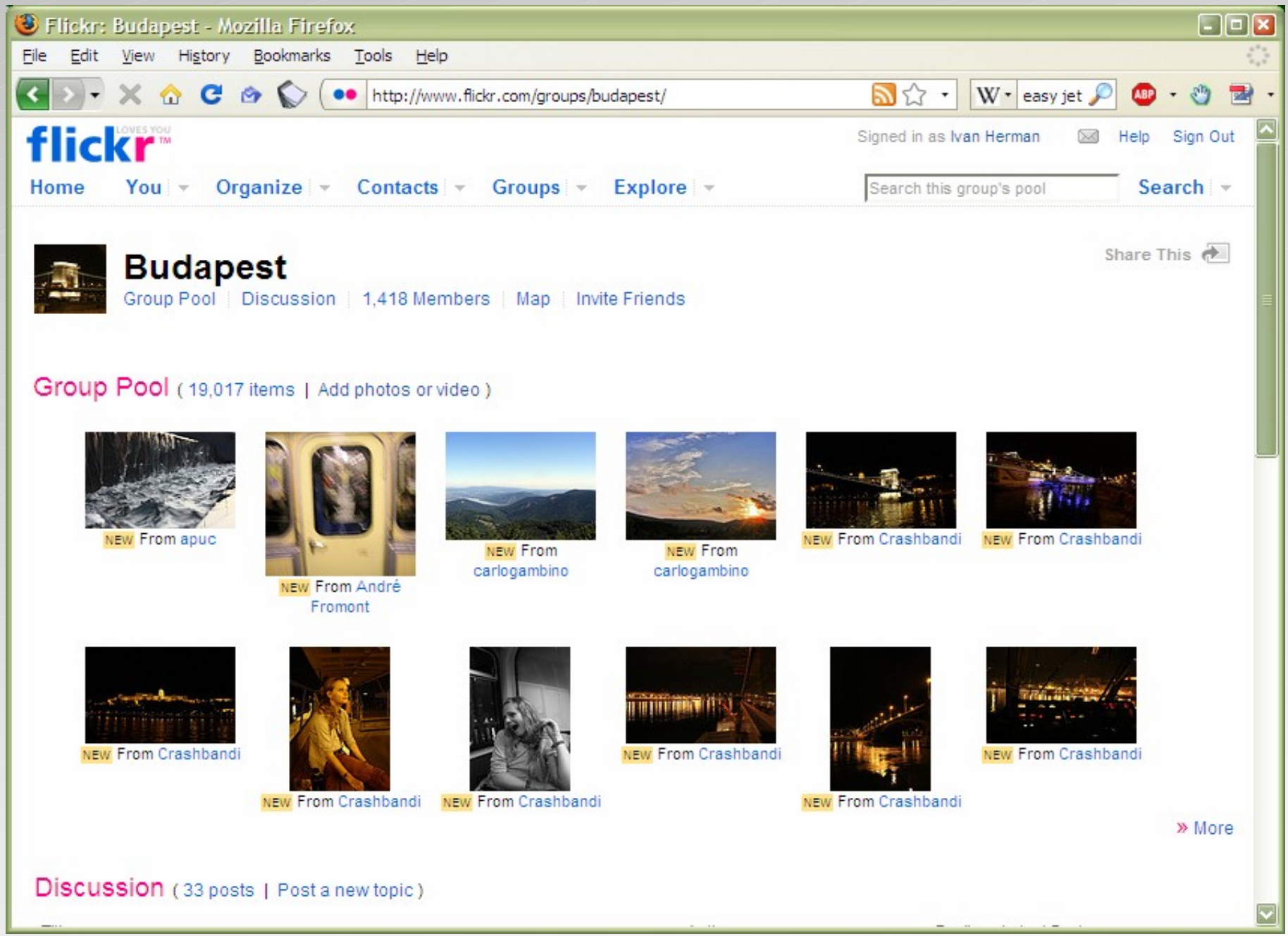


… or this one

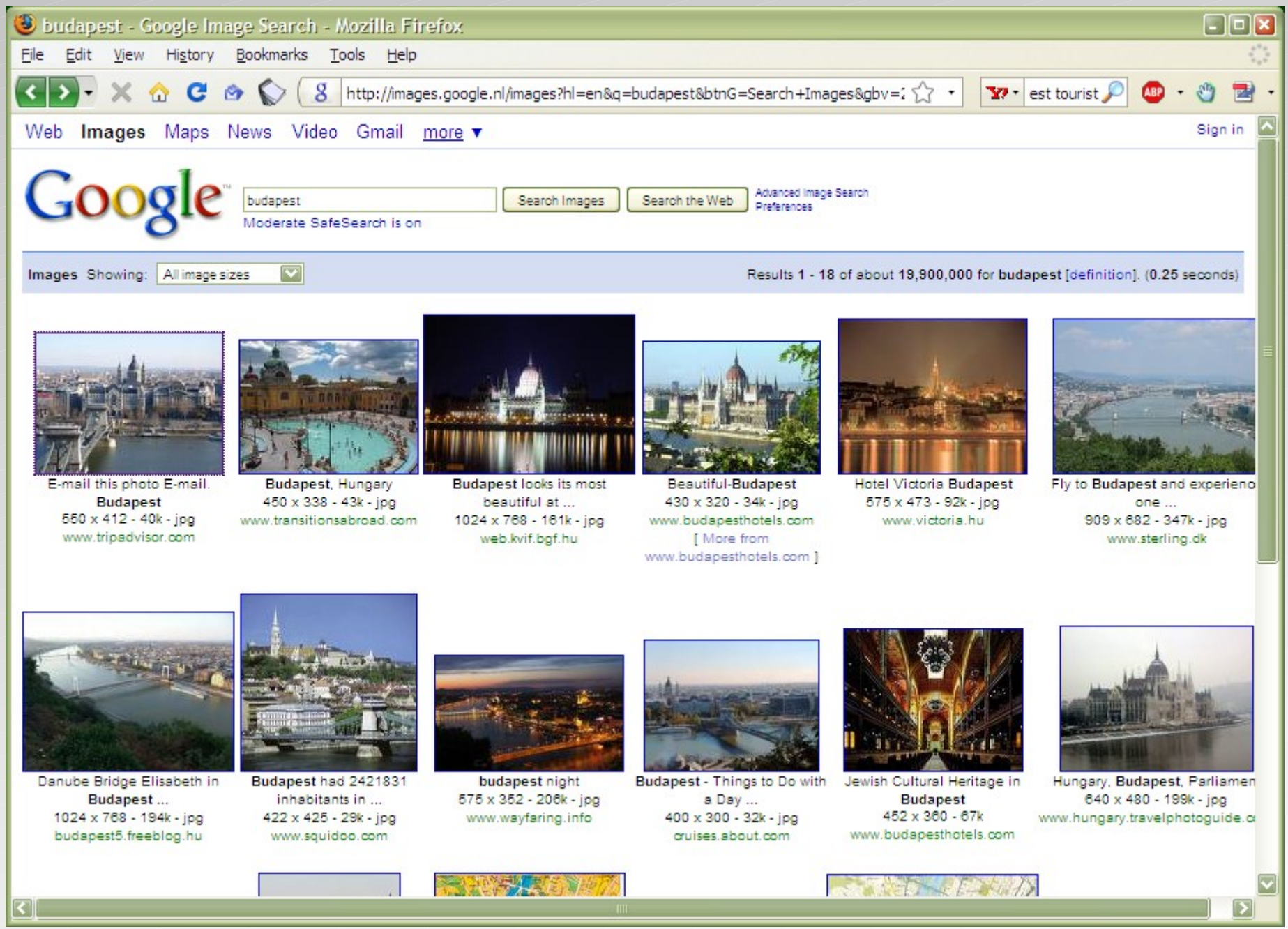


You may want to know something about Budapest; look for some photographs…

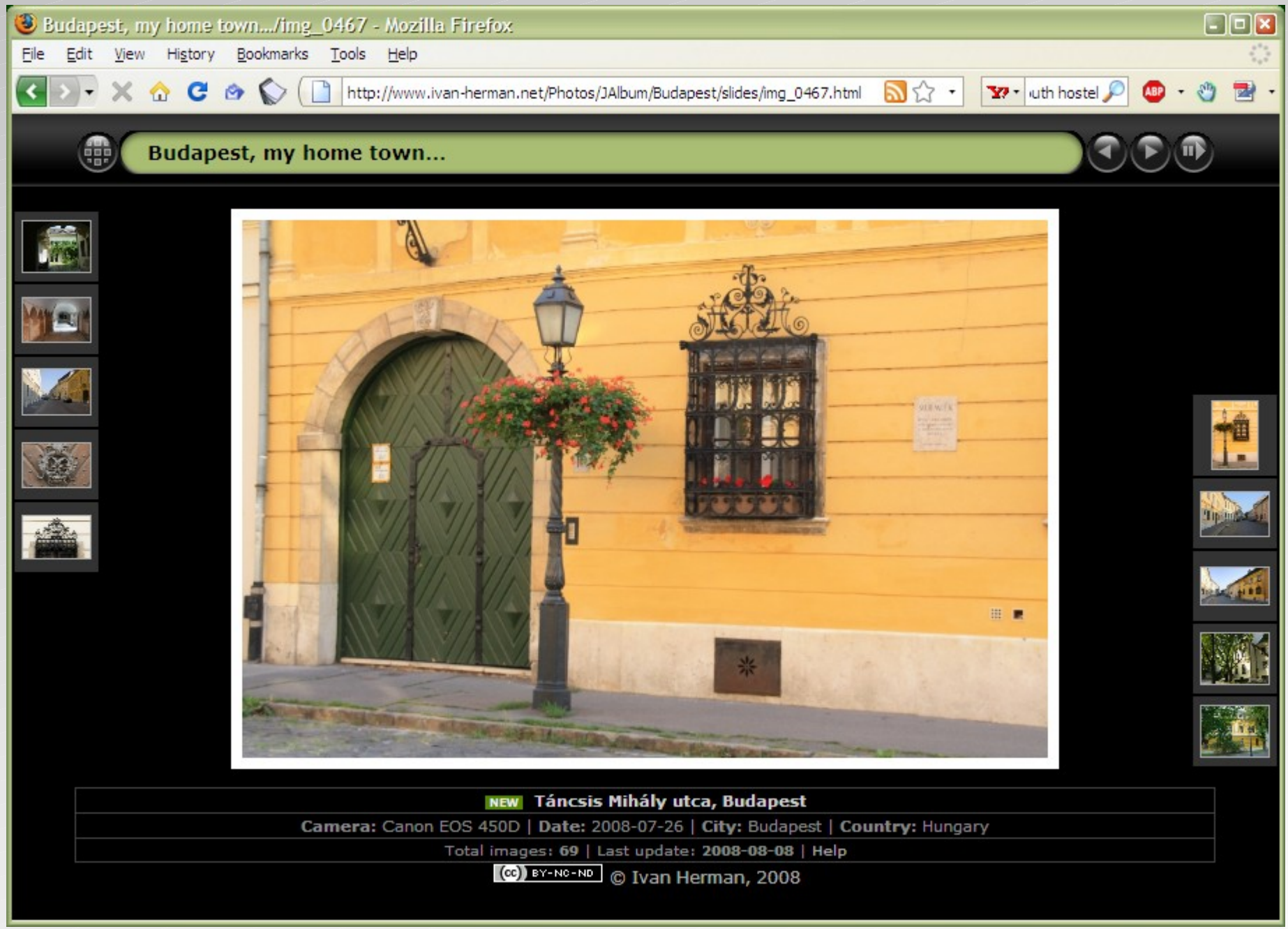
… on flickr …



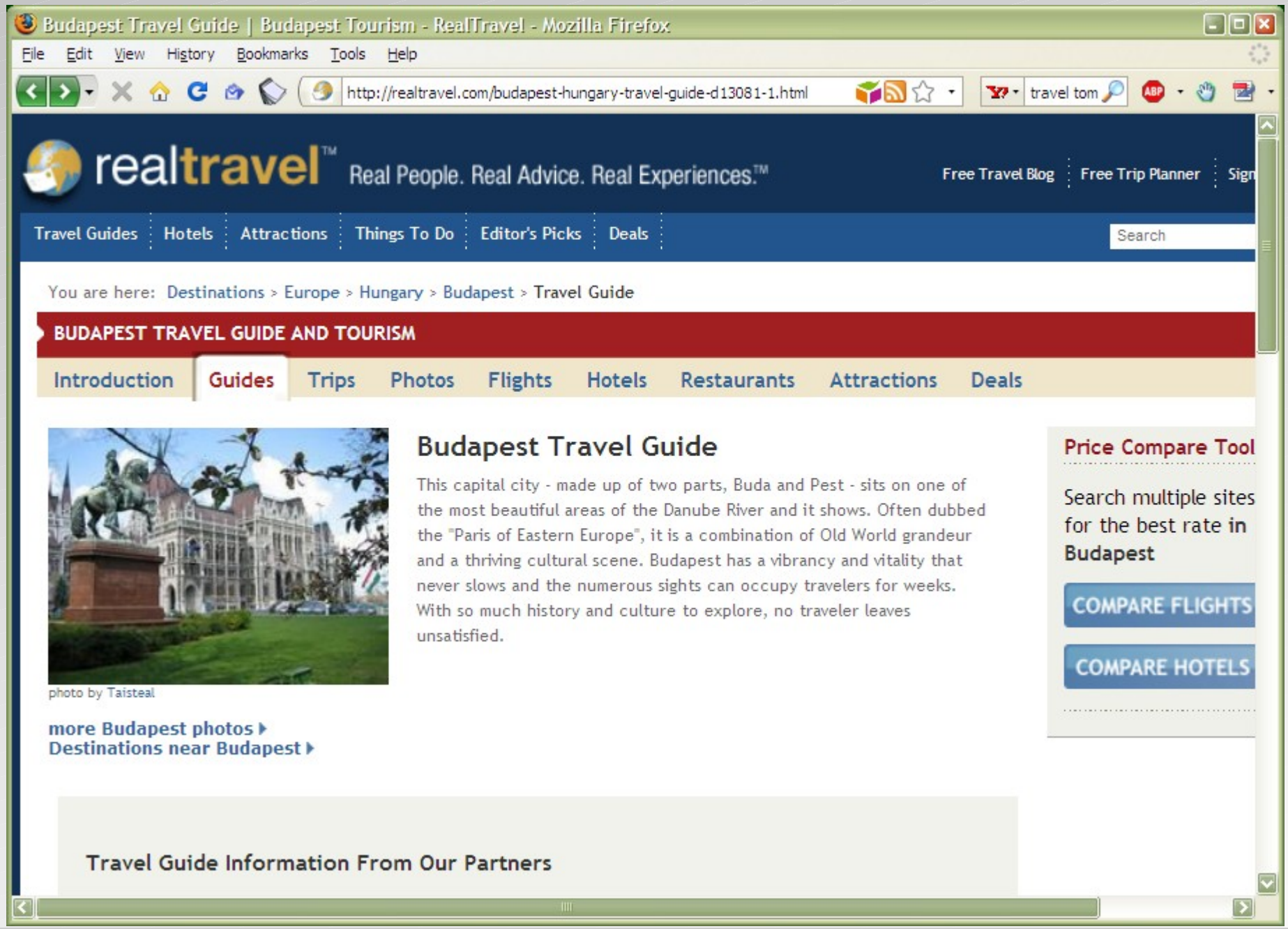
… on Google …



… or you can look at mine



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



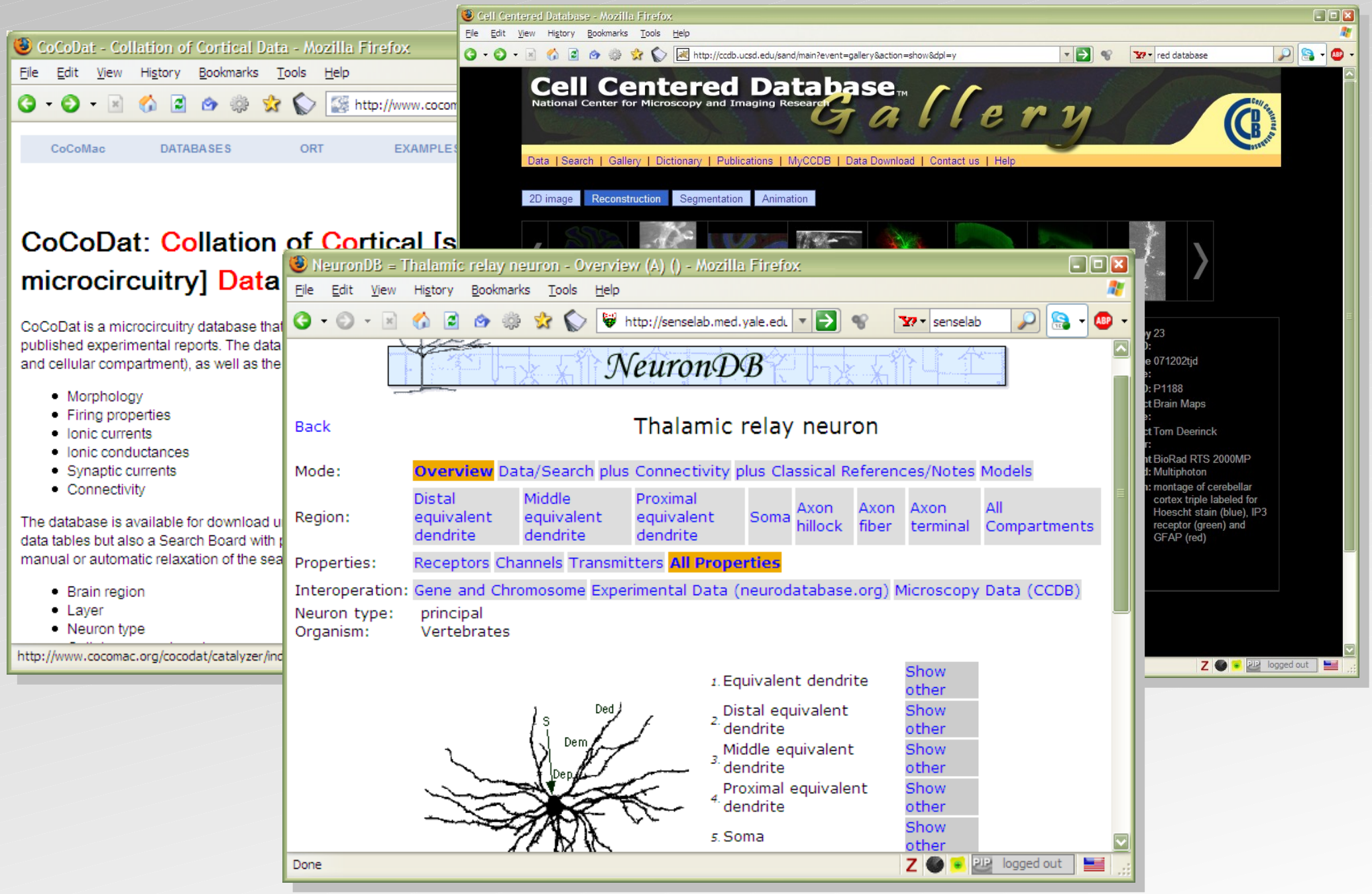
What happened here?

- 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!

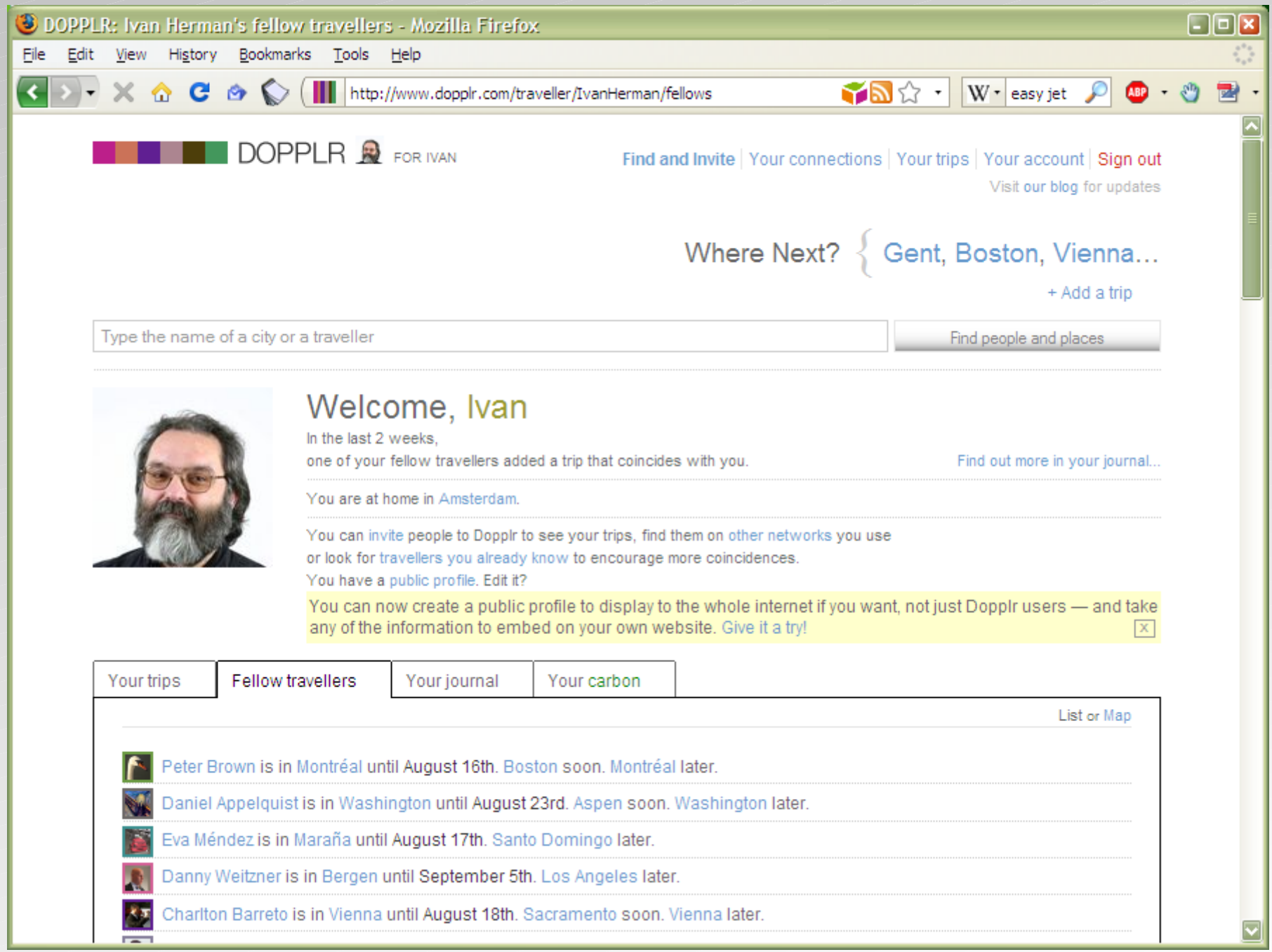
Here is another example…



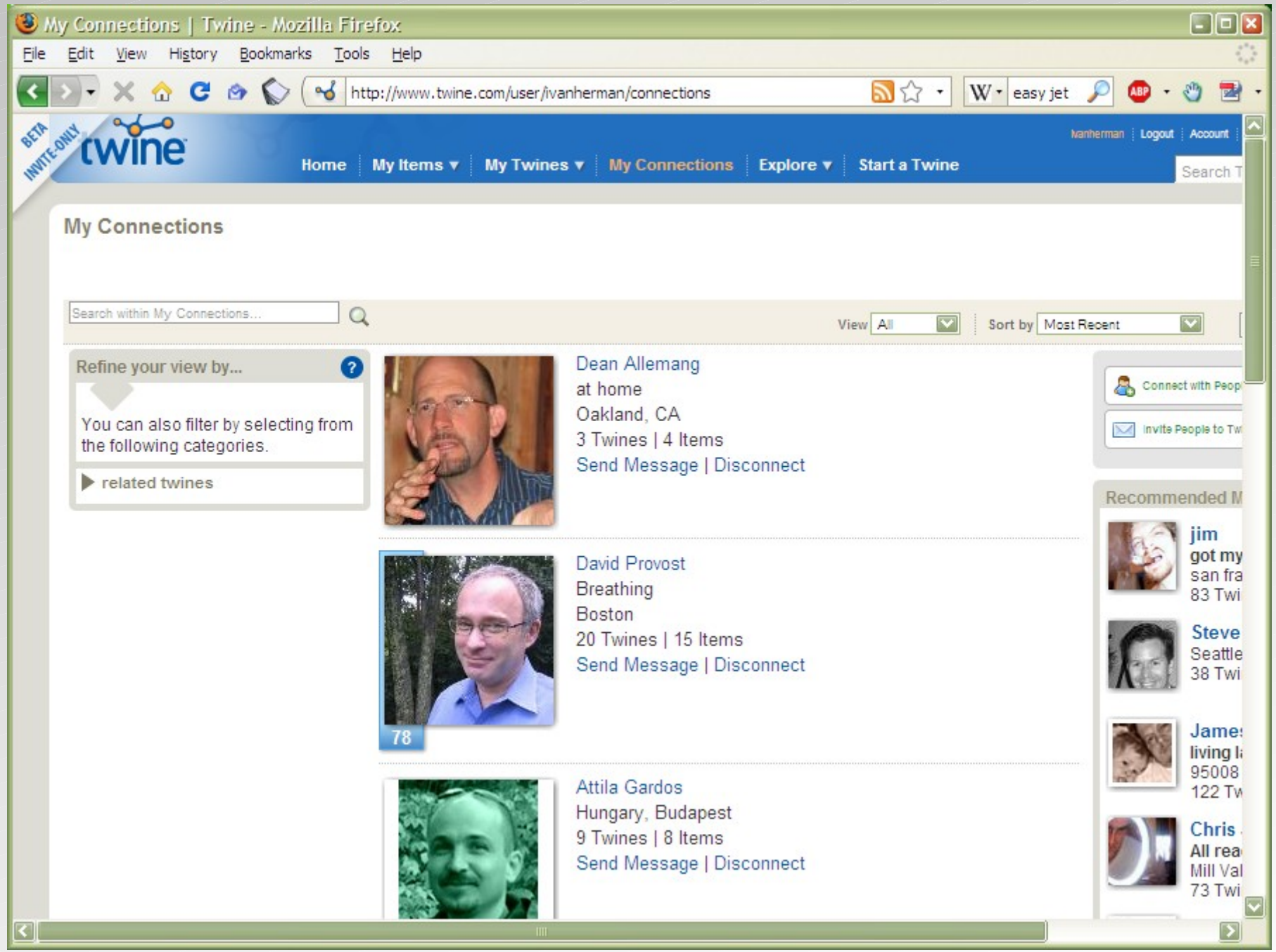
- Companies may have to hire a person to answer questions based on those (public!) databases!

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

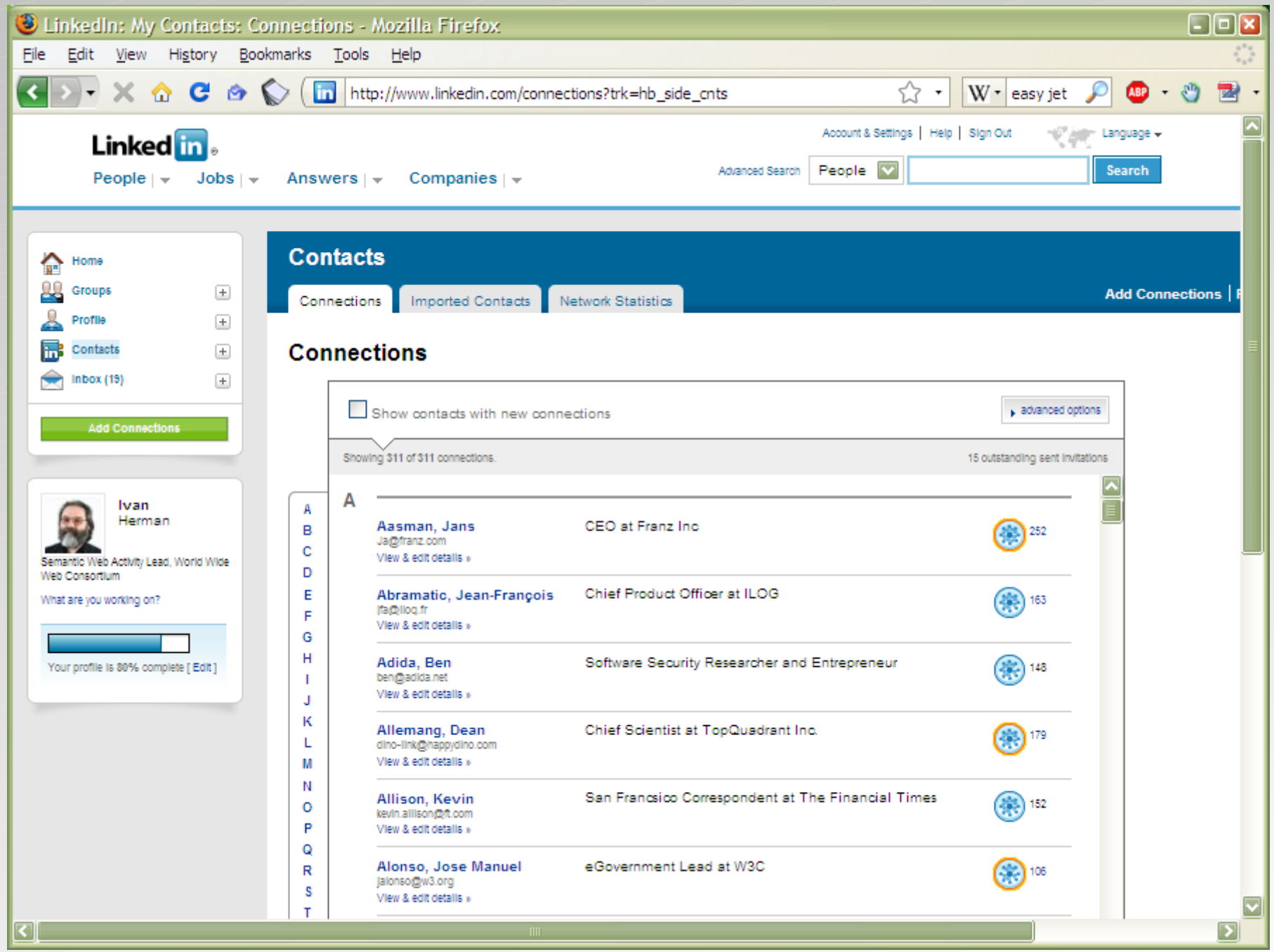
… Dopplr,



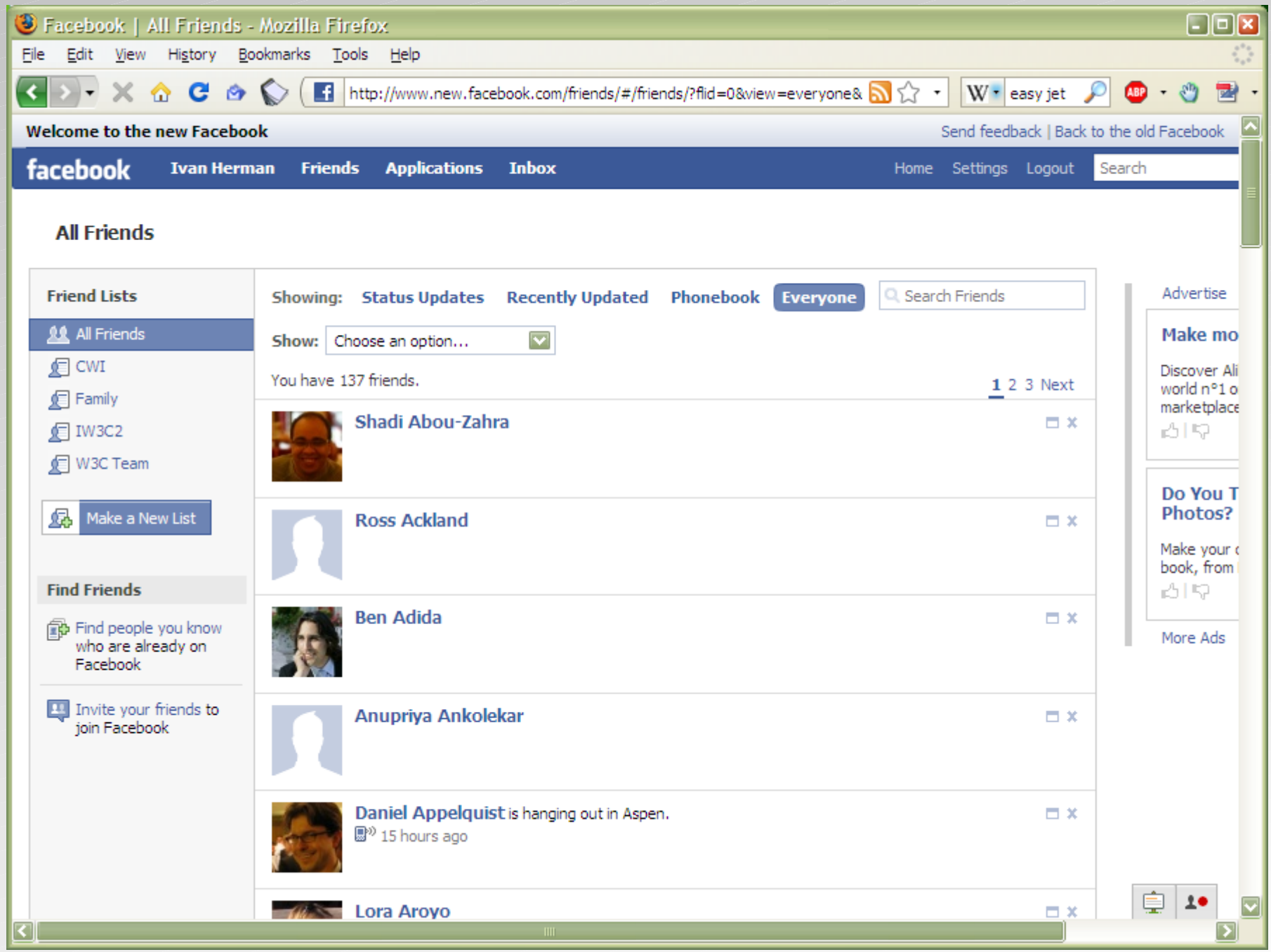
… Twine,



… 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?

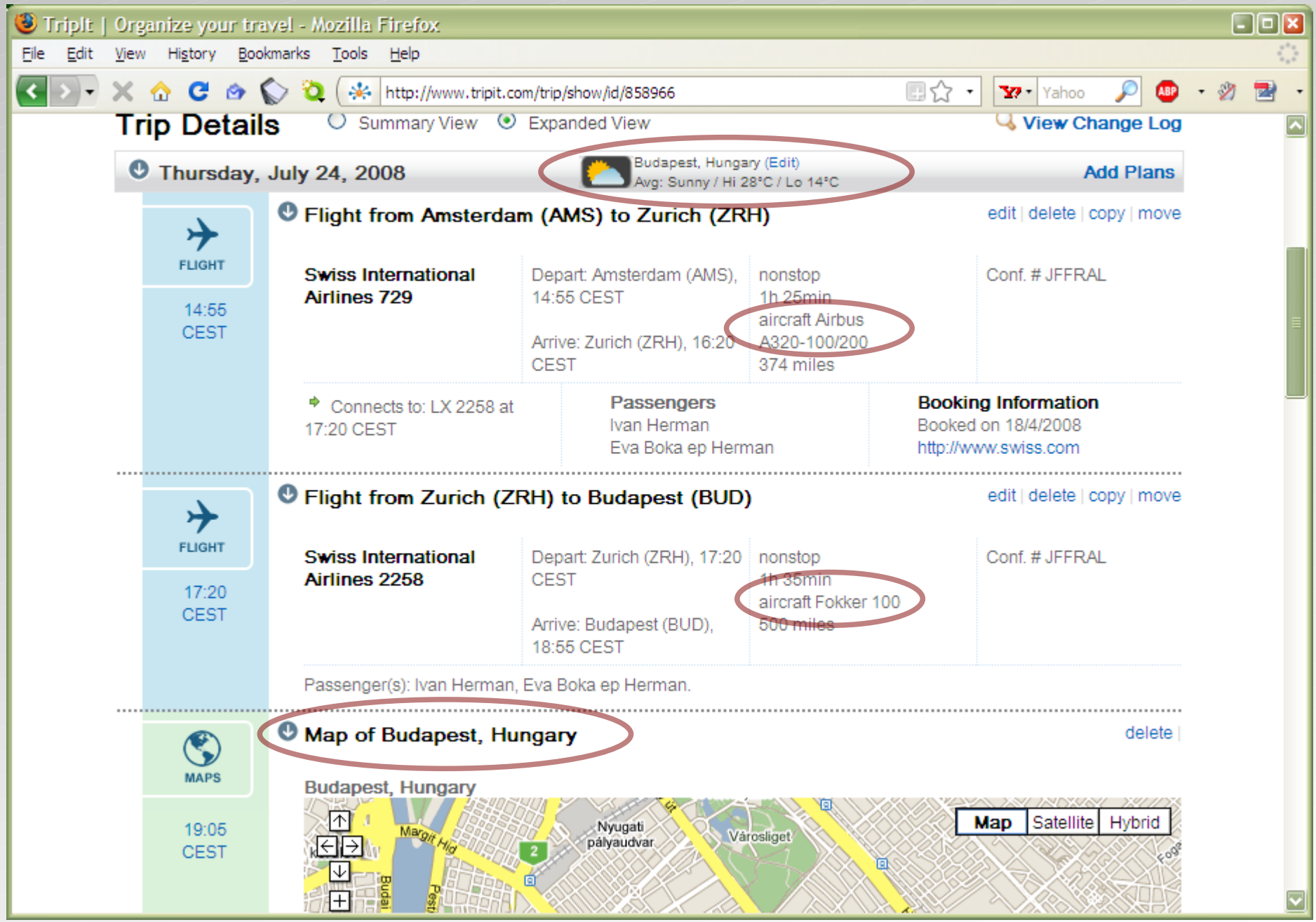
- Use the data on the Web the same way as we do with documents:
  - be able to link to data (independently of their presentation)
  - use that data the way I want (present it, mine it, etc)
  - agents, programs, scripts, etc, should be able to *interpret* part of that data
- *"The bane of my existence is to do things that my computer could do for me"* — Dan Connolly

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?

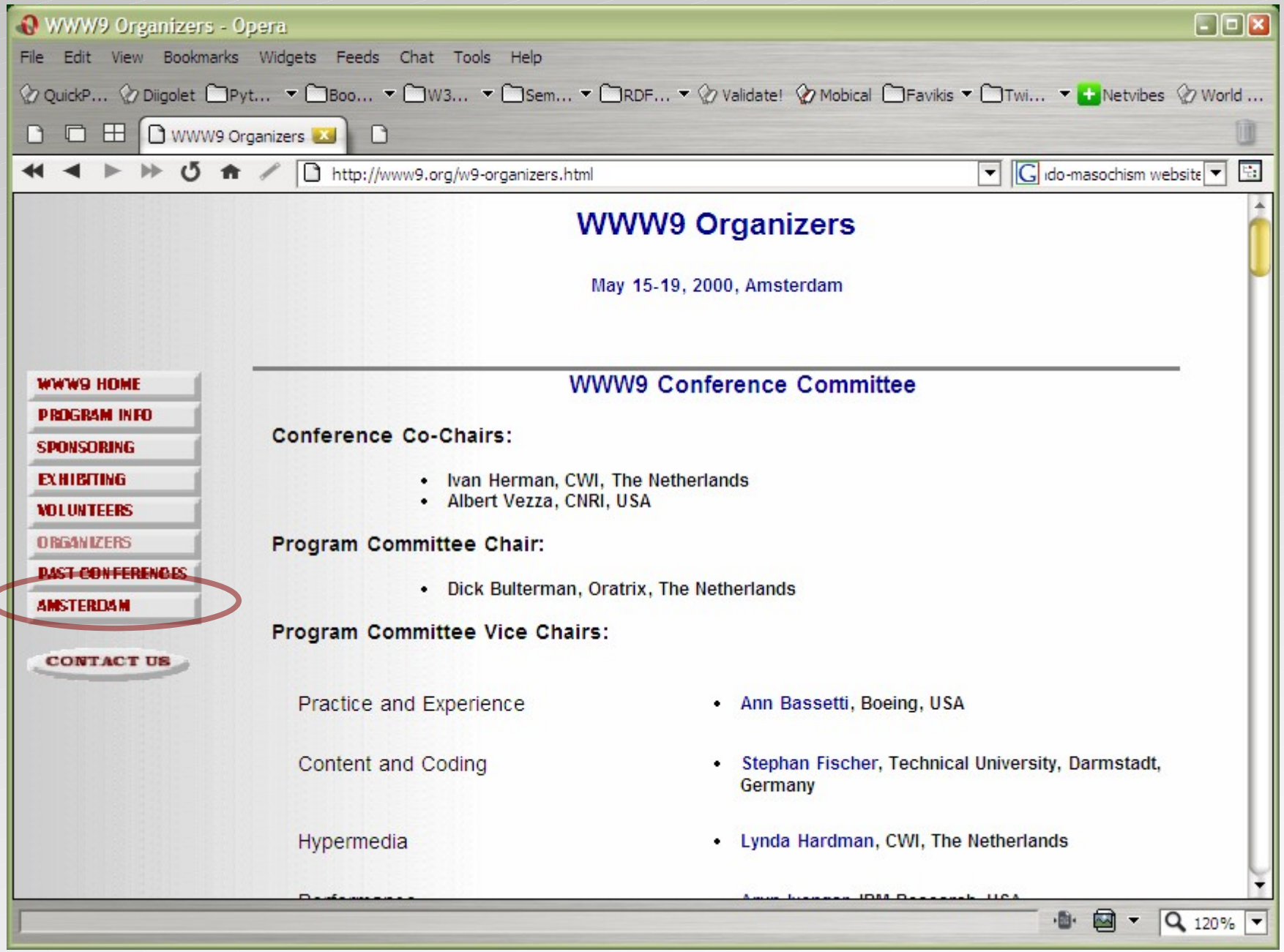
- What makes the current (document) Web work?
  - people create different documents
  - they give an address to it (ie, a URI) and make it accessible to others on the Web

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

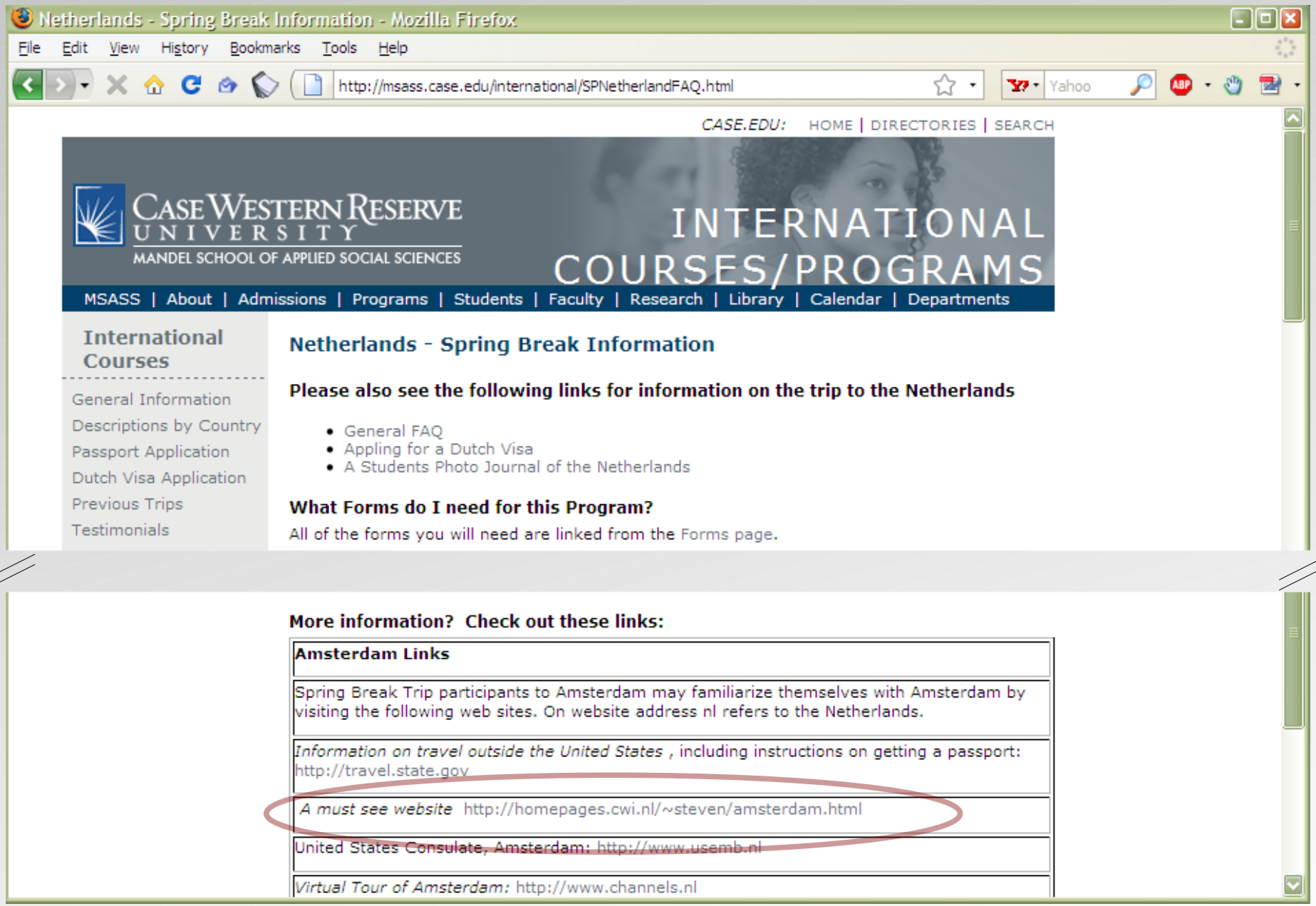
Then some magic happens…

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

This could be expected…



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

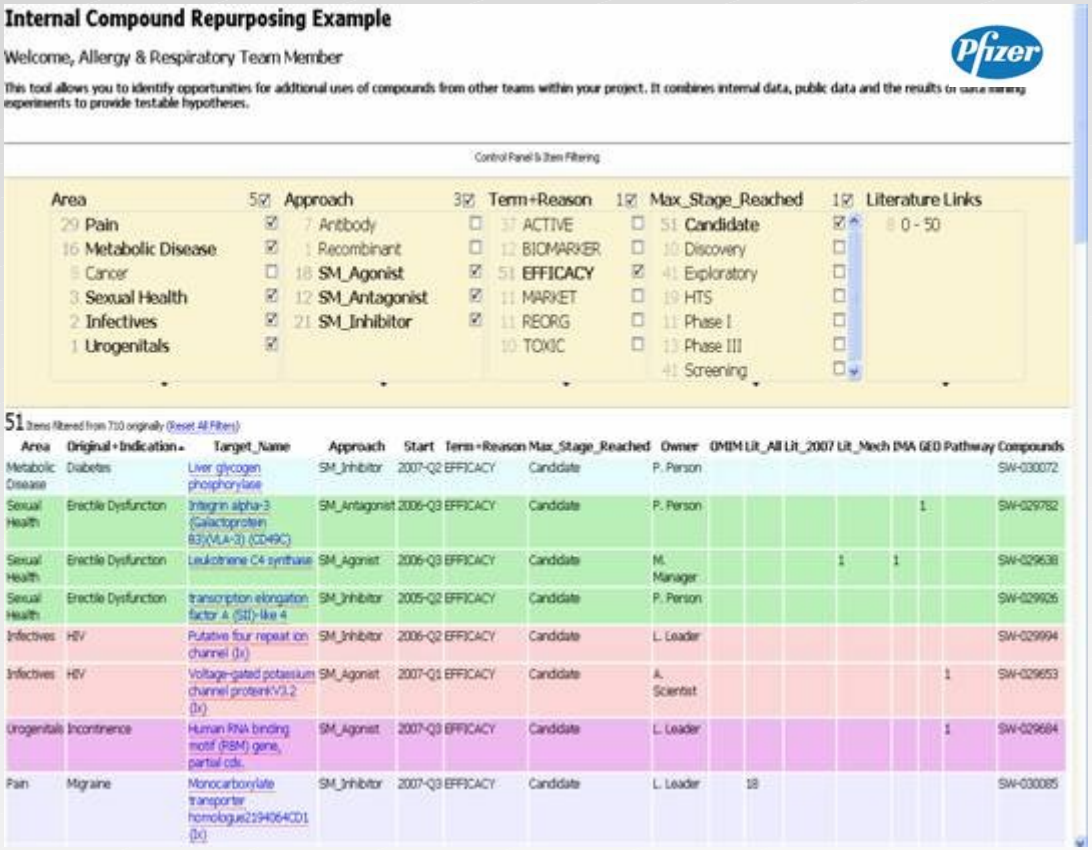


What would that mean for a Web of Data?

- 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

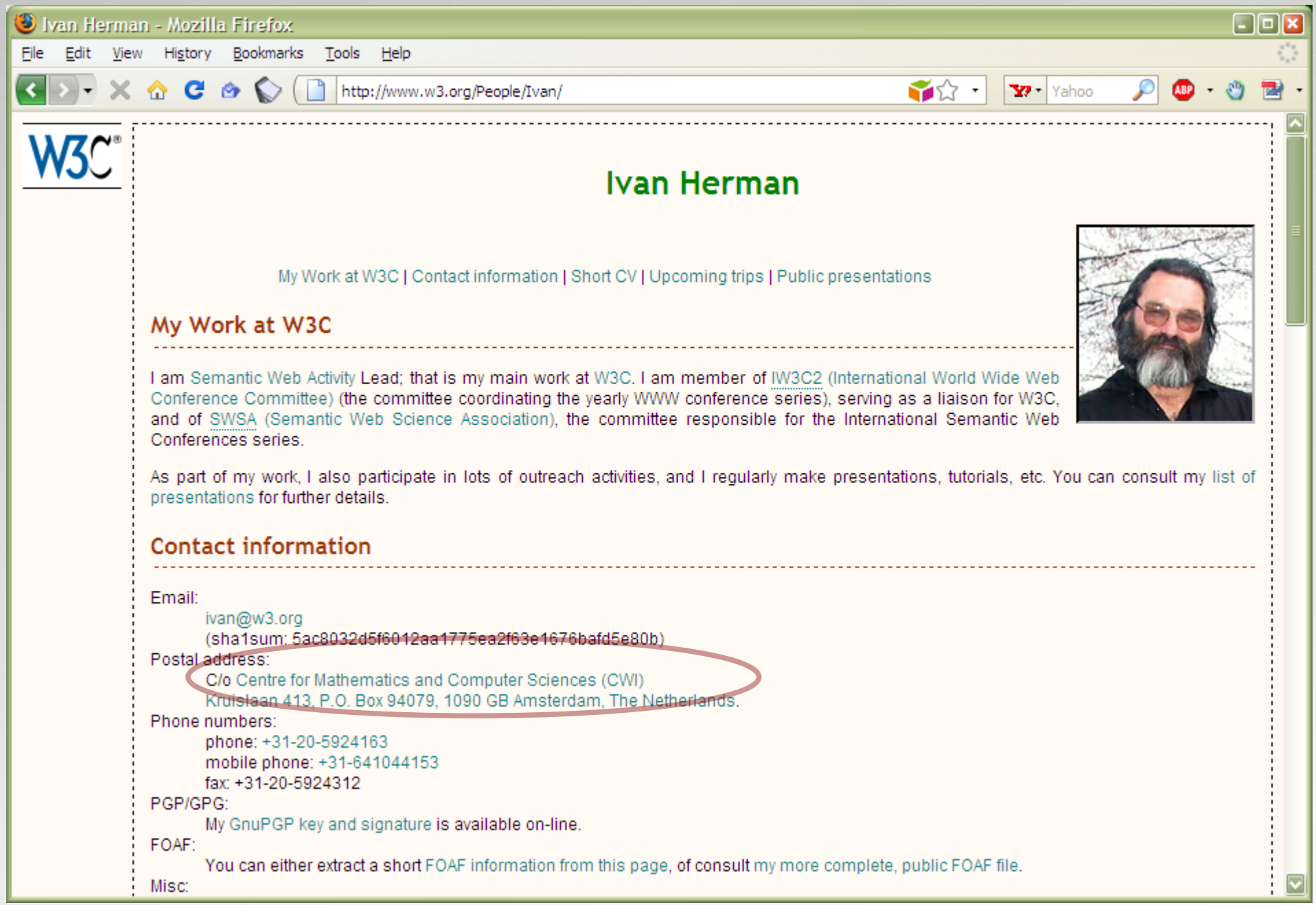


Courtesy of Nigel Wilkinson, Lee Harland, Pfizer Ltd, Melliyal Annamalai, Oracle (SWEO Case Study)

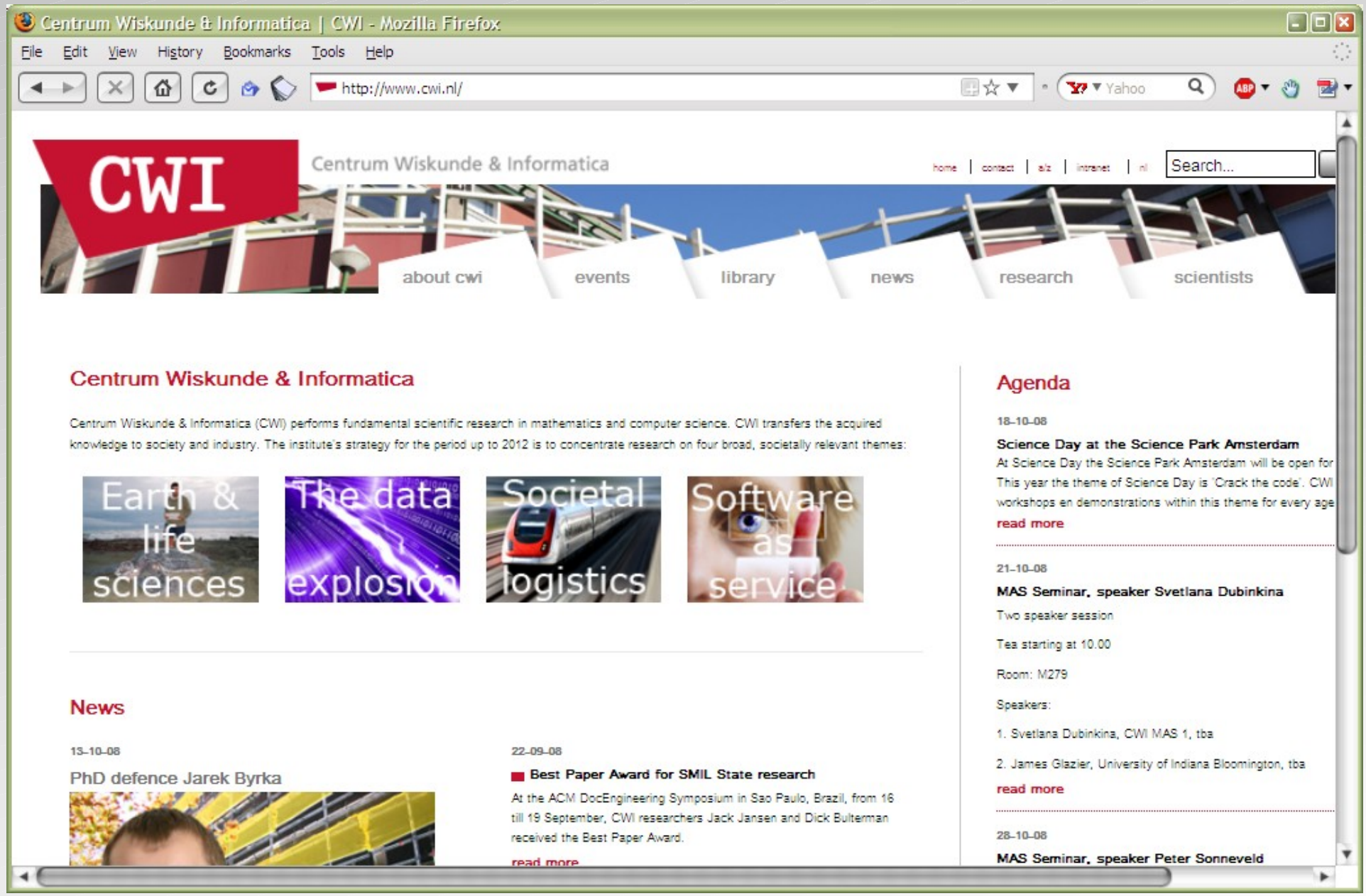
But it is a little bit more complicated

- 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 (realize that it is a research institute in Amsterdam)
- On a Web of Data, something is missing; machines can make no sense of the link alone

- New lesson learned:
  - extra information (“label”) must be added to a link: “this links to my institution”
  - 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!

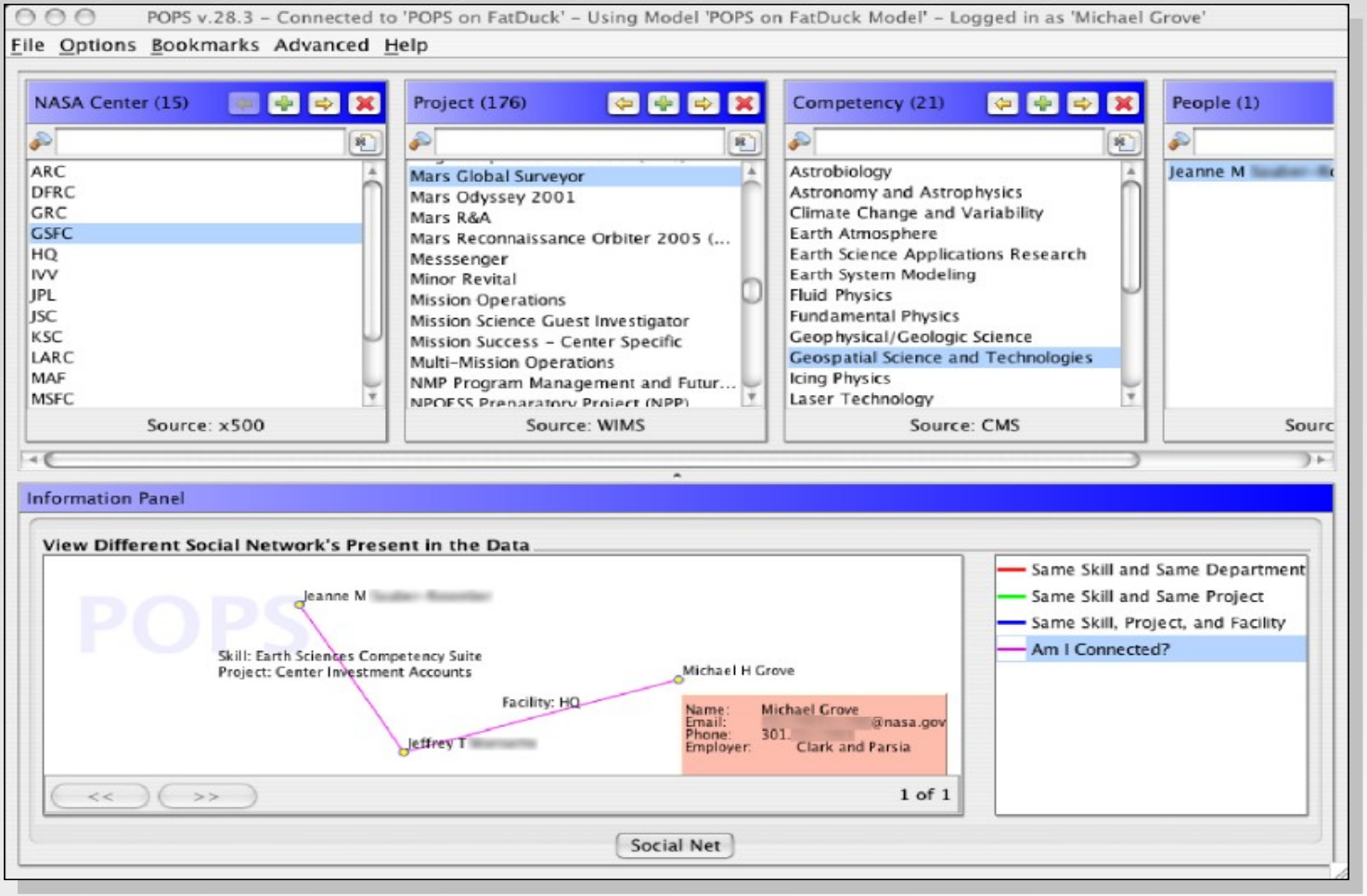
Example: find the right experts at NASA

- 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!

Michael Grove, Clark & Parsia, LLC, and Andrew Schain, NASA, (SWEO Case Study)

Example: find the right experts at NASA

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



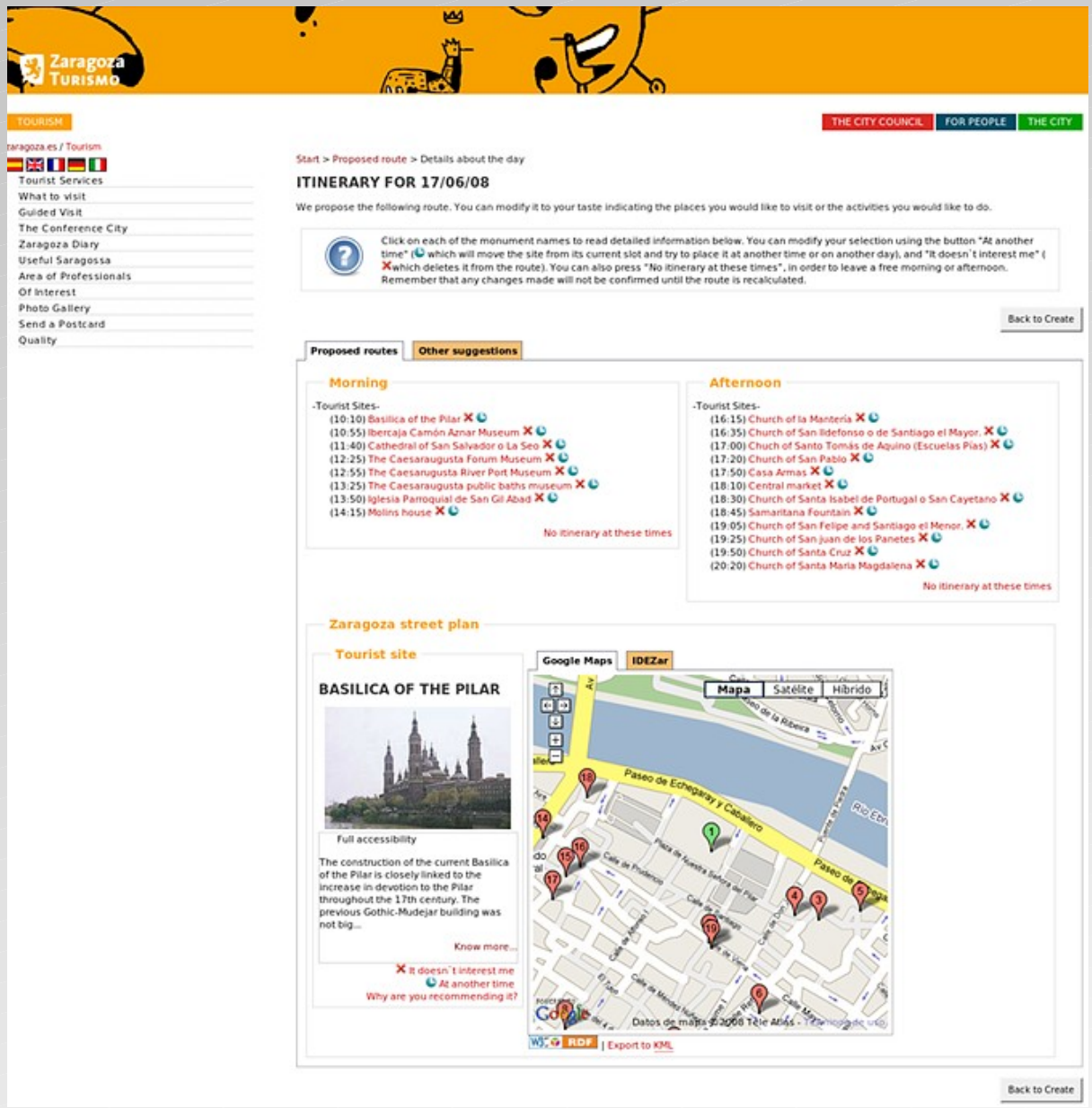
Michael Grove, Clark & Parsia, LLC, and Andrew Schain, NASA, (SWEO Case Study)

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 understand the “labels” and draw some conclusions from that info
  - the “classification” of the terms can become *very* complex for specific knowledge areas: this is where ontologies, thesauri, vocabularies, etc, enter the game…

Example: eTourism in Zaragoza



- Provide personalized itinerary service
- Integration of different databases in Zaragoza (using targeted ontologies)

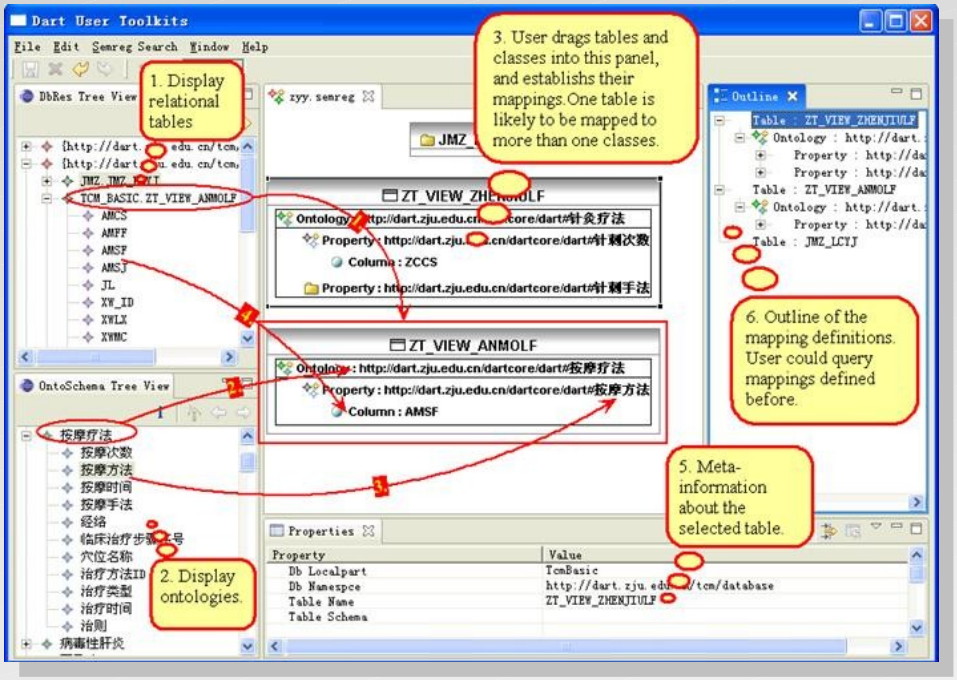
Courtesy of Jesús Fernández, Municipality of Zaragoza, and Antonio Campos, CTIC (SWEO Use Case)

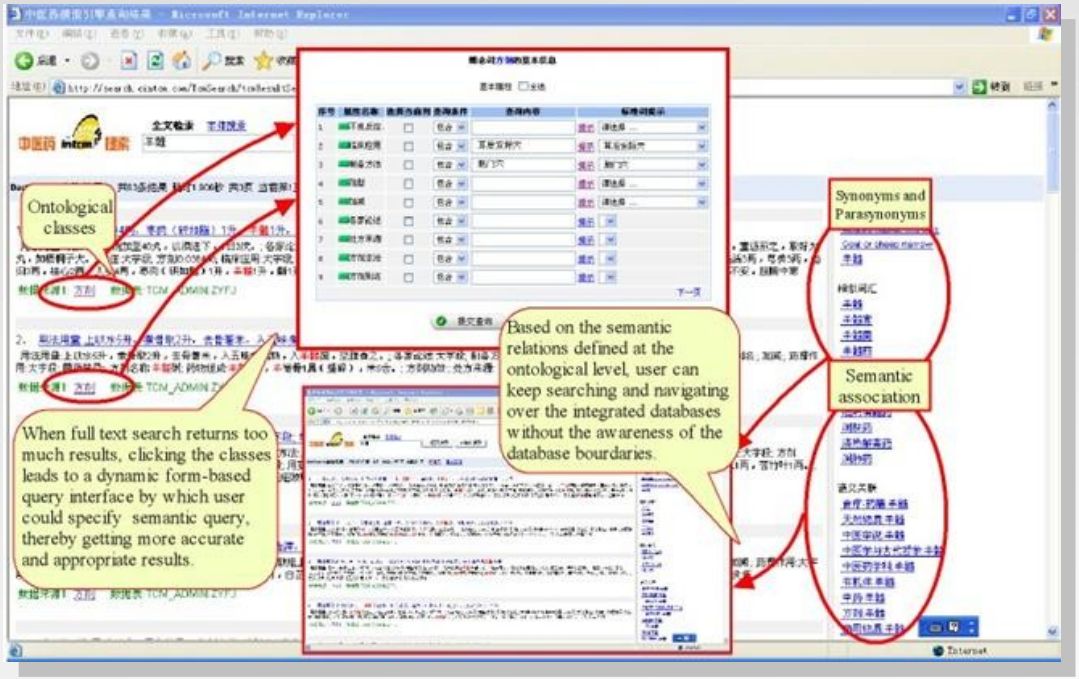
Wait! Does it mean that I have to convert all my data in some way?

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

Example: integrate knowledge for Chinese Medicine

- Integration of a large number of TCM databases
  - around 80 databases, around 200,000 records each
- Uses specialized ontologies
- Queries are converted on-the-fly to the databases





Courtesy of Huajun Chen, Zhejiang University, (SWEO Case Study)

In the end…

- More an more data should be “published” on the Web
  - this can lead to the “network effect” on data
- Different communities will develop their own vocabularies
- New breed of applications
  - “mashups on steroids”
  - better representation and usage of community knowledge
  - new customization possibilities
  - …