Introduction to the Semantic Web (through an Example…)

Ivan Herman, W3C

(Last updated: 31 October 2008)

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 infos may come from different sites
  - searches in different digital libraries
  - etc.
- Again, humans combine these information easily
  - even if different terminology's 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?
  - …

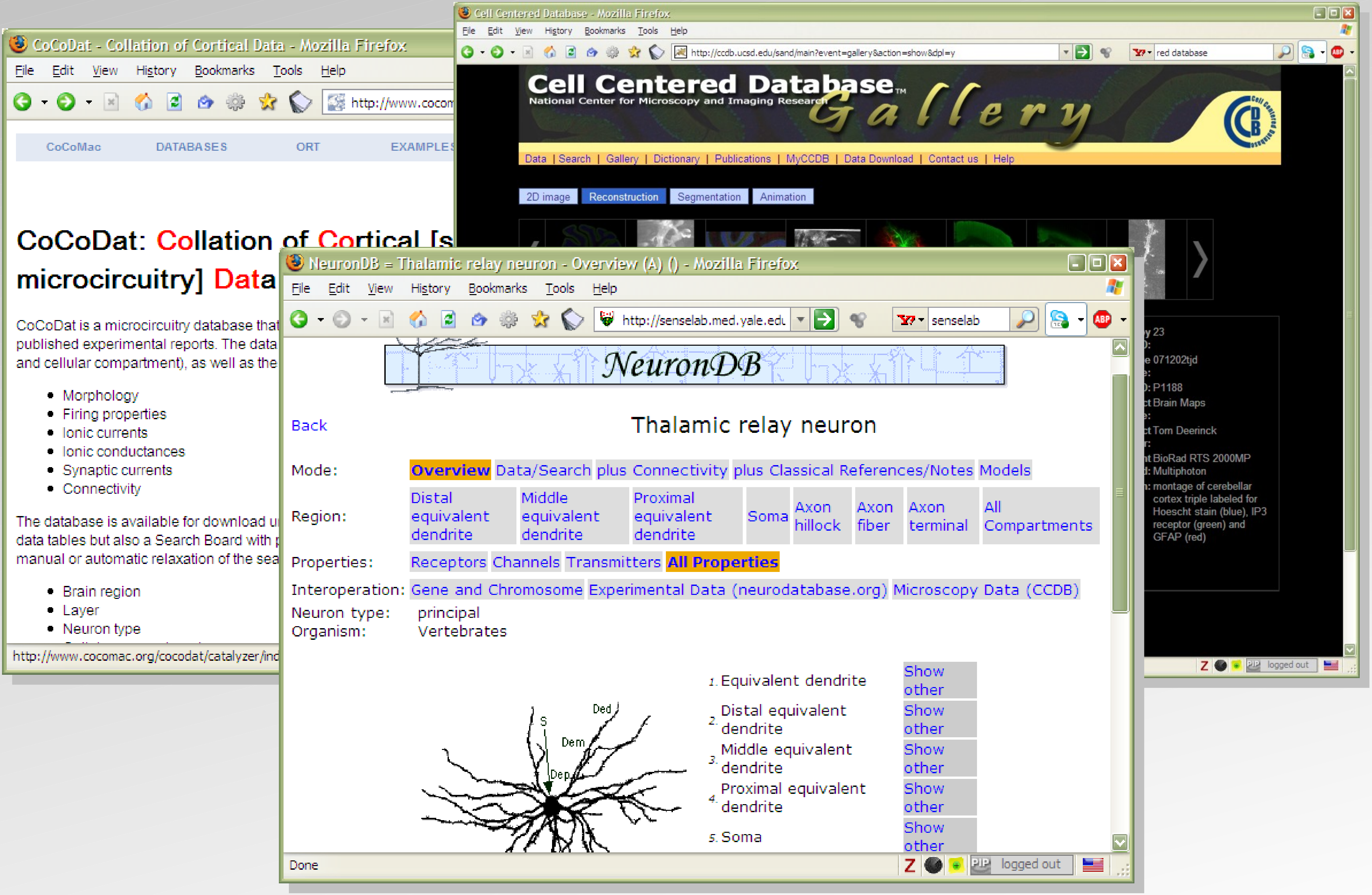
Example: automatic airline reservation

- 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…


CoCoDat - Collation of Cortical Data - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.cocom
CoCoMac DATABASES ORT EXAMPLES

CoCoDat: Collation of Cortical [s
microcircuitry] Data

CoCoDat is a microcircuitry database that
published experimental reports. The data
and cellular compartment), as well as the

- Morphology
- Firing properties
- Ionic currents
- Ionic conductances
- Synaptic currents
- Connectivity

The database is available for download u
data tables but also a Search Board with
manual or automatic relaxation of the sea

- Brain region
- Layer
- Neuron type

http://www.cocomac.org/cocodat/catalyzer/ind
Cell Centered Database - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://ccdb.ucsd.edu/sand/main?event=gallery&action=show&dpl=y
red database
Cell Centered Database™
National Center for Microscopy and Imaging Research
Gallery
Data | Search | Gallery | Dictionary | Publications | MyCCDB | Data Download | Contact us | Help
2D image
Reconstruction
Segmentation
Animation

y 23
:
e 071202tjd
e:
): P1188
ct Brain Maps
e:
ct Tom Deerinck
r:
nt BioRad RTS 2000MP
d: Multiphoton
n: montage of cerebellar cortex triple labeled for Hoescht stain (blue), IP3 receptor (green) and GFAP (red)

logged out
NeuronDB = Thalamic relay neuron - Overview (A) () - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://senselab.med.yale.edu
senselab
NeuronDB

Back

Thalamic relay neuron

Mode: **Overview** Data/Search plus Connectivity plus Classical References/Notes Models

Region: Distal equivalent dendrite | Middle equivalent dendrite | Proximal equivalent dendrite | Soma | Axon hillock | Axon fiber | Axon terminal | All Compartments

Properties: Receptors Channels Transmitters **All Properties**

Interoperation: Gene and Chromosome Experimental Data (neurodatabase.org) Microscopy Data (CCDB)

Neuron type: principal
Organism: Vertebrates

1. Equivalent dendrite Show other
2. Distal equivalent dendrite Show other
3. Middle equivalent dendrite Show other
4. Proximal equivalent dendrite Show other
5. Soma Show other

Example: Social Networks

- Social sites are everywhere these days (LinkedIn, Facebook, Dopplr, Digg, Plexo, Zyb, …)
- Data is not interchangeable: how many times did you have to add your contacts? 😕
- Applications should be able to get to those data via standard means
  - there are, of course, privacy issues…

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: change of address & the authorities

- It means change of address at “official” places
    - so you could still get the right official mails for official notices, tax information, certificates, etc.
- … but you never know if you notified the right local, regional, national, etc, authorities
    - ie, you still get some mail from some agency at your old address ☹
- It should be possible to change the address in one official place only
    - the administration should be smart enough to propagate the changes
    - this means that various authorities should be able to merge their data…

Example: “smart” portal

- Various types of “portals” are created (for a journal on-line, for a specific area of knowledge, for specific communities, etc)
- The portals may:
  - integrate lots of different data sources
  - may have access to specialized domain knowledge
- Goal is to provide a better local access, search on the integrated data, reveal new relationships among the data

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

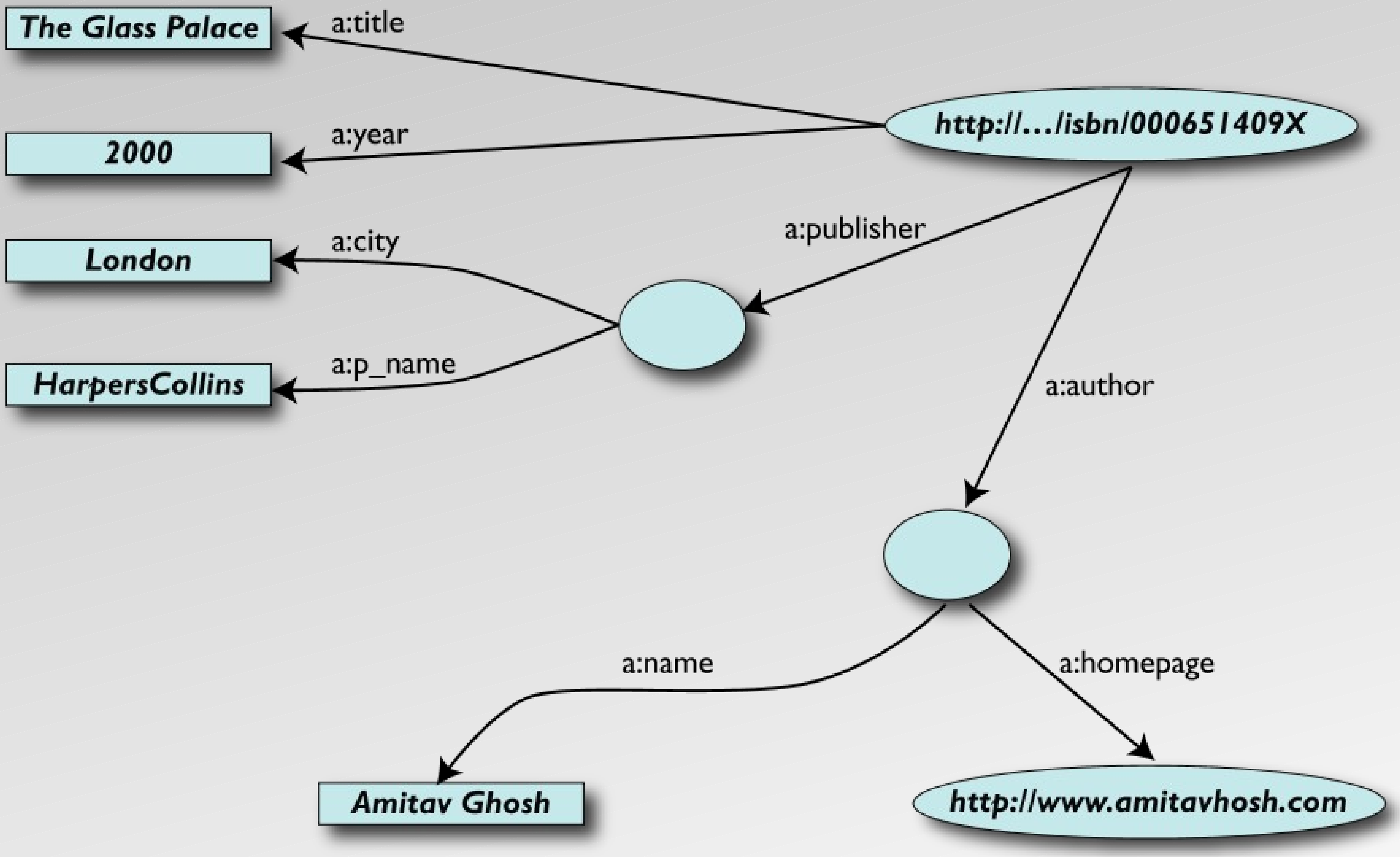
A *simplified* bookstore data (dataset “A”)

ID	Author	Title	Publisher	Year
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	Harpers Collins	London

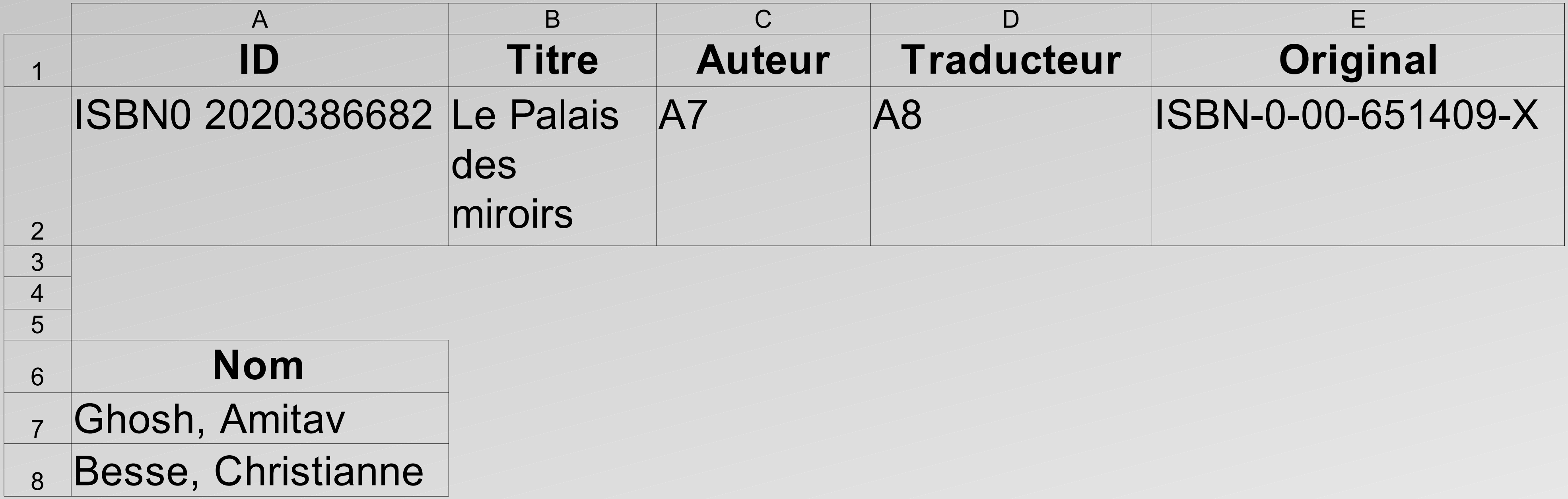
1st: export your data as a set of *relations*



Some notes on the exporting the data

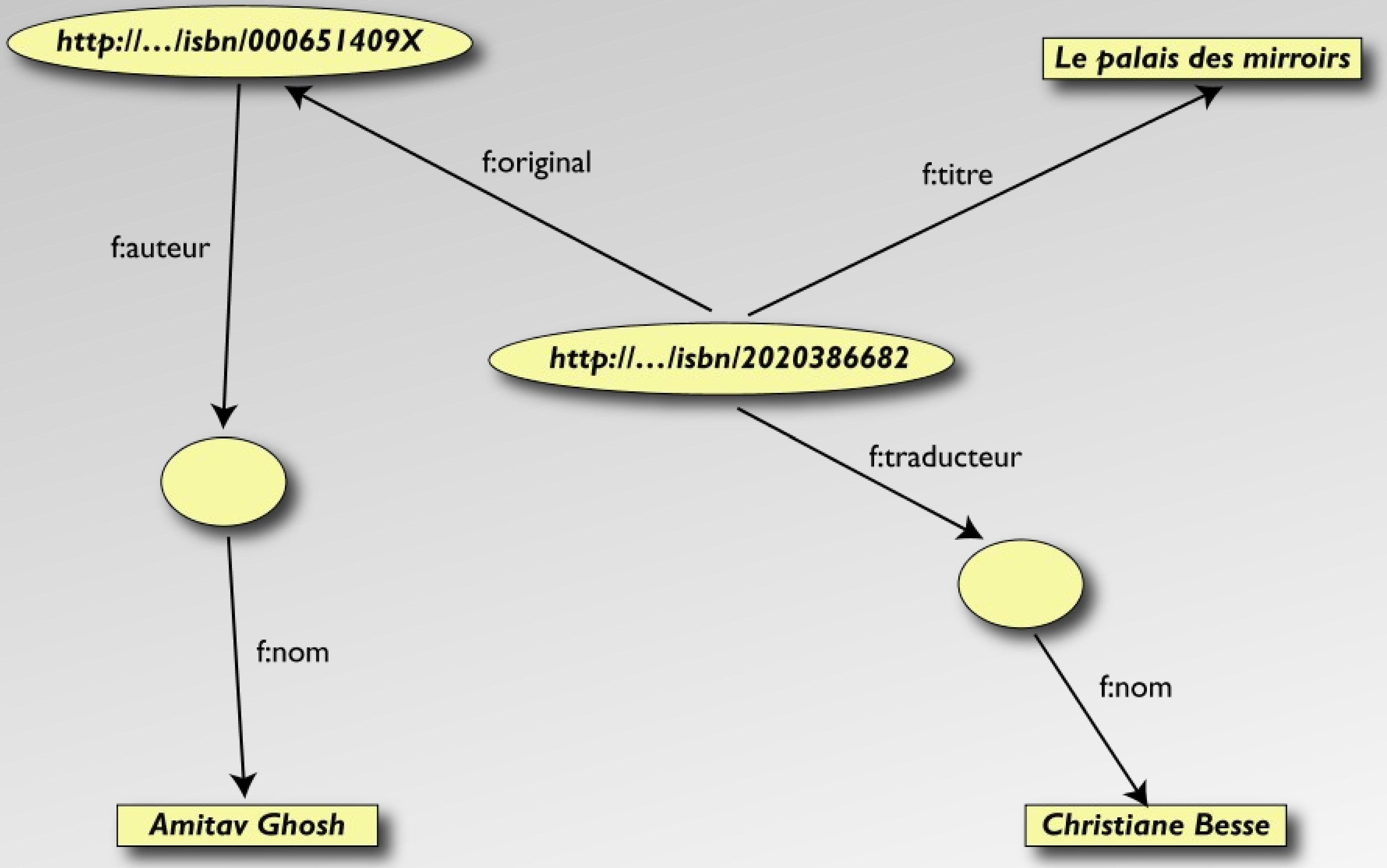
- Relations form a graph
  - the nodes refer to the “real” data or contain some literal
  - how the graph is represented in machine is immaterial for now
- Data export does *not* necessarily mean physical conversion of the data
  - relations can be generated on-the-fly at query time
    - via SQL “bridges”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

Another bookstore data (dataset “F”)

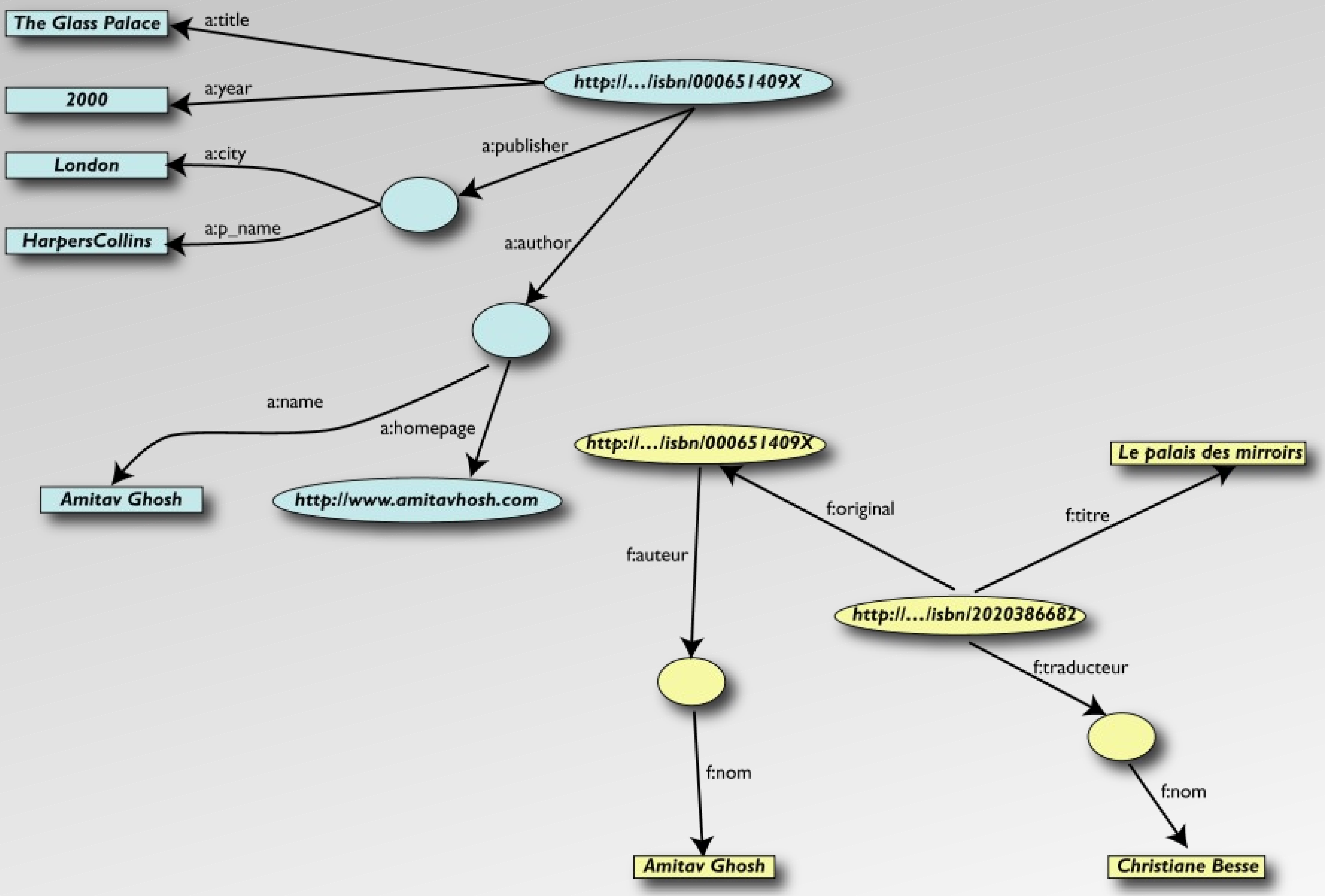


	A	B	C	D	E
1	**ID**	**Titre**	**Auteur**	**Traducteur**	**Original**
2	ISBN0 2020386682	Le Palais des miroirs	A7	A8	ISBN-0-00-651409-X
3					
4					
5					
6	**Nom**				
7	Ghosh, Amitav				
8	Besse, Christianne				

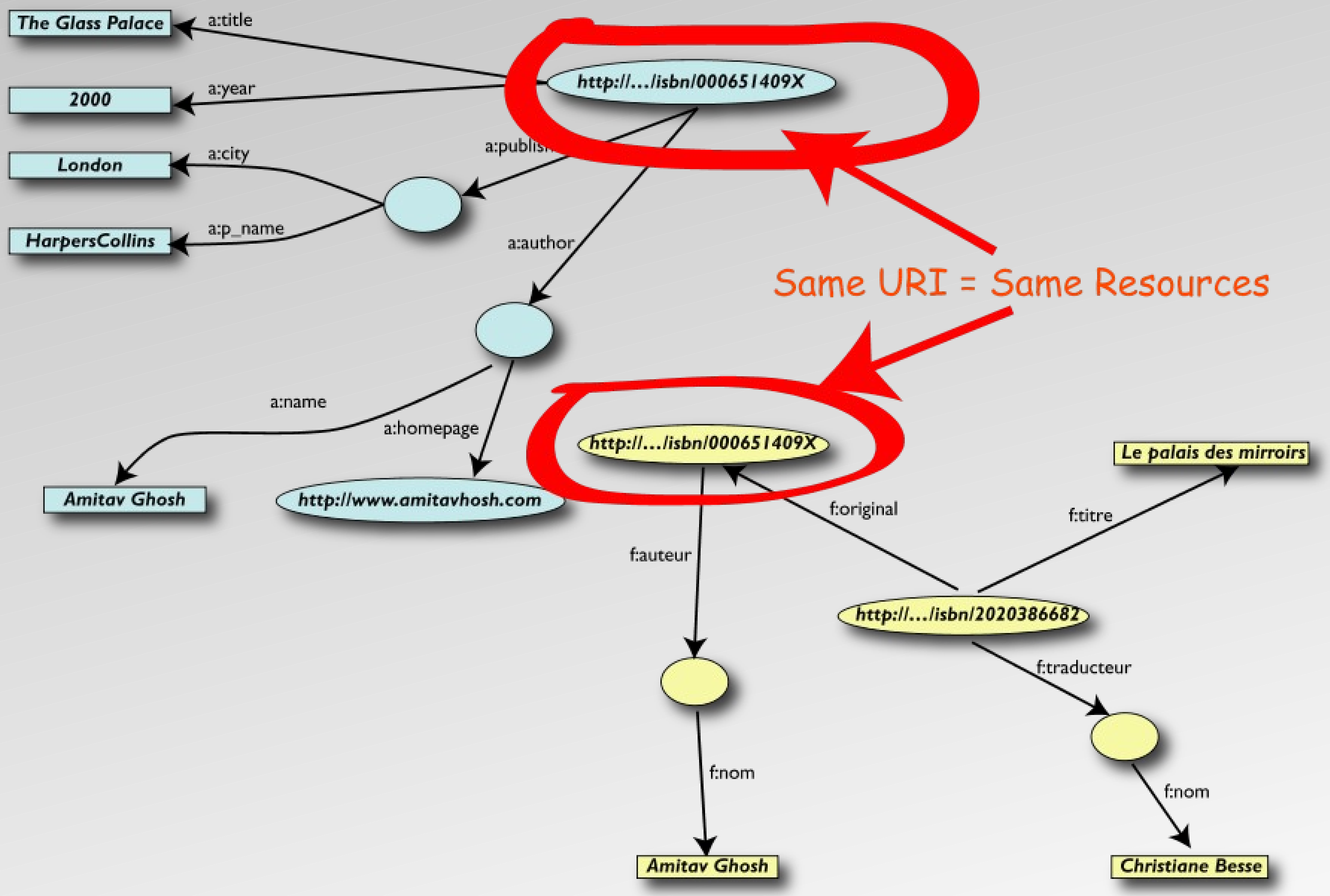
2nd: export your second set of data



3rd: start merging your data

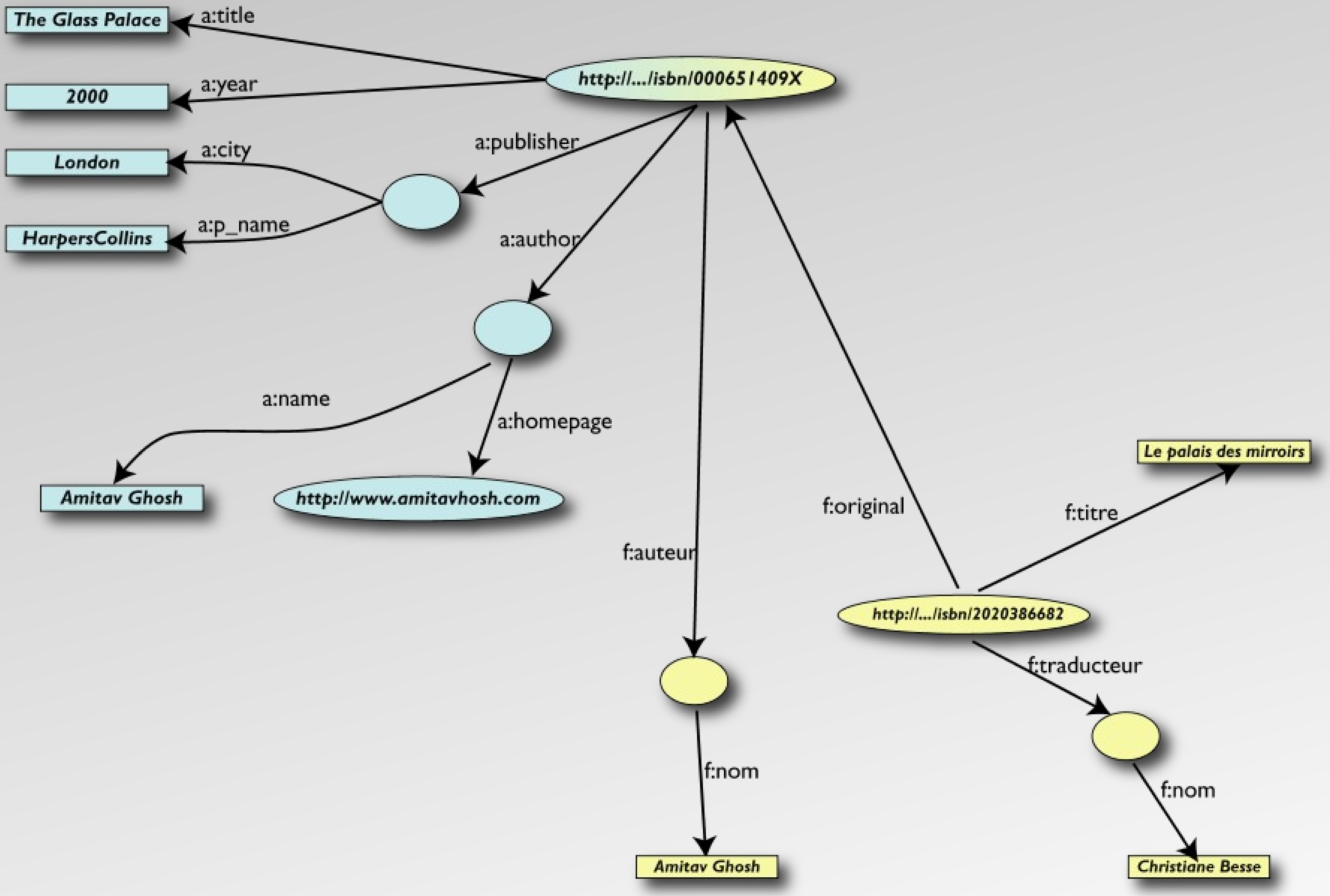


3rd: start merging your data (cont.)



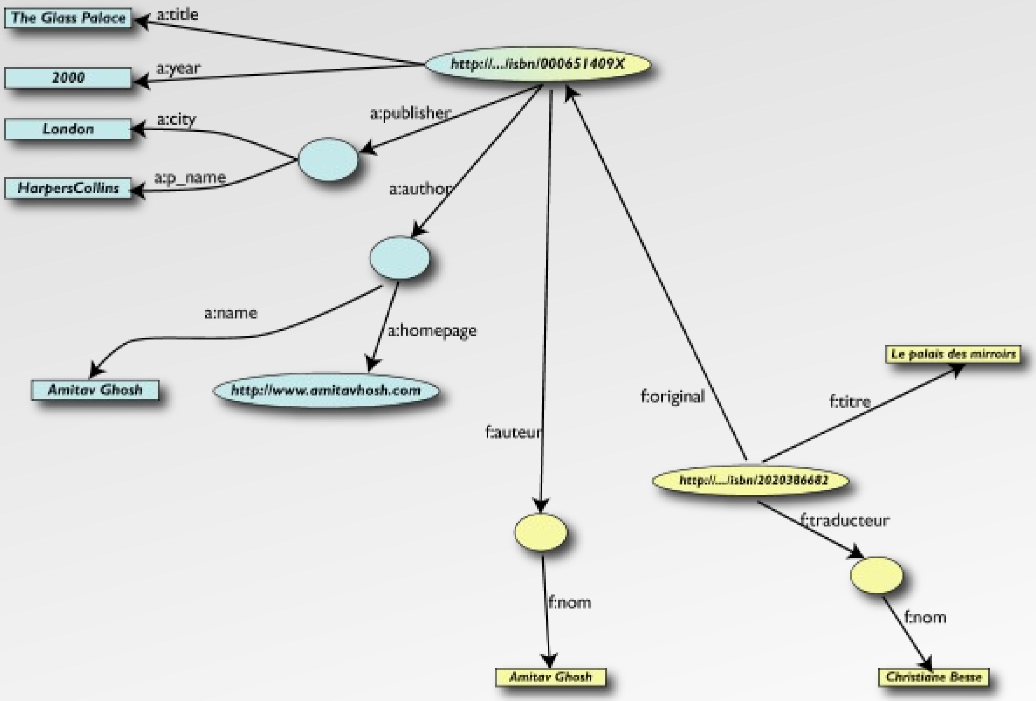
Same URI = Same Resources

3rd: merge identical resources



Start making queries…

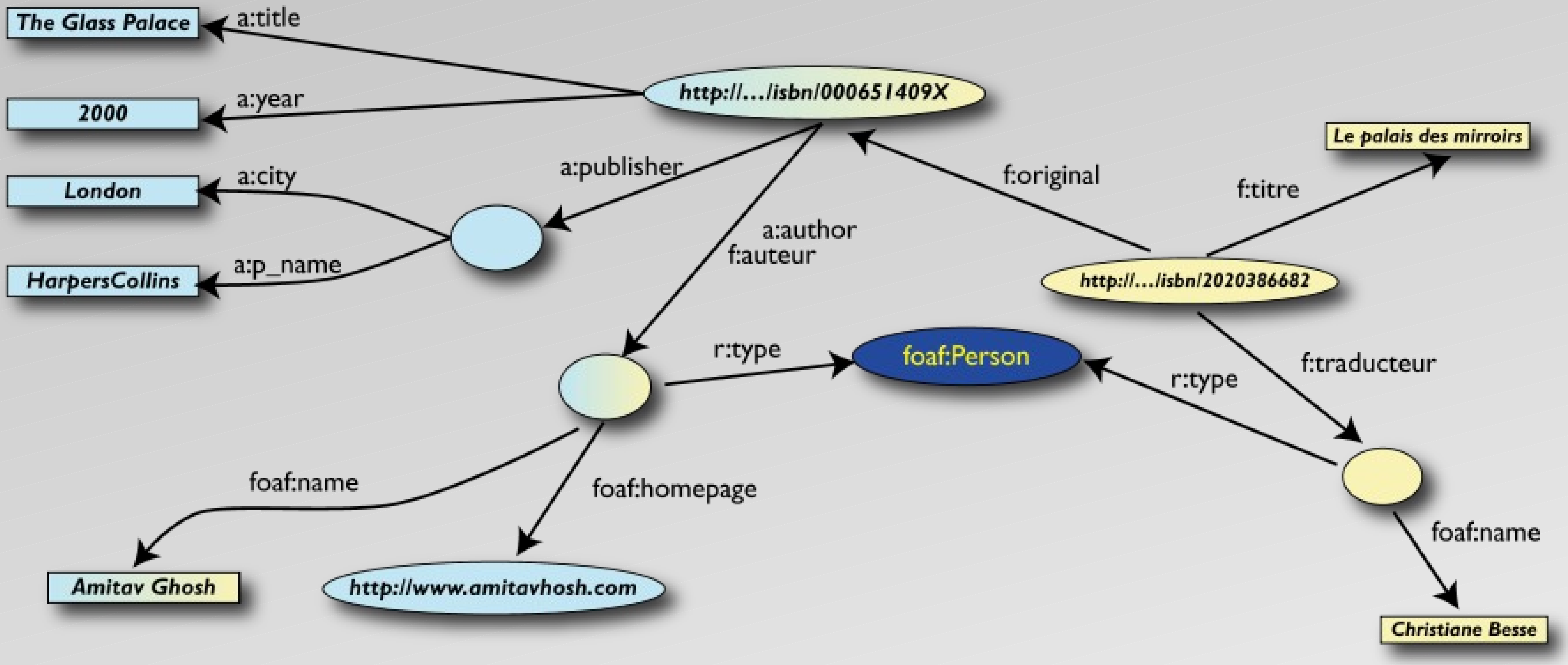
- User of data “F” can now ask queries like:
  - “give me the title of the original”
    - well, … « donnes-moi le titre de l’original »
- This information is not in the dataset “F”…
- …but can be retrieved by merging with dataset “A”!

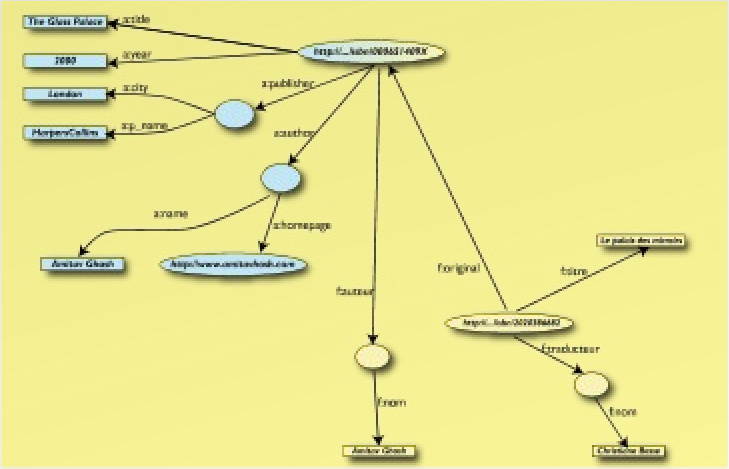


However, more can be achieved…

- 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

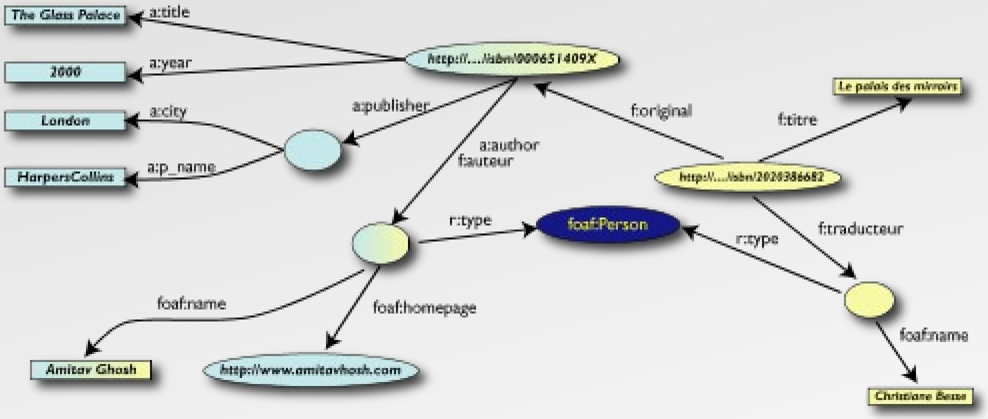
3rd revisited: use the extra knowledge





Start making richer queries!

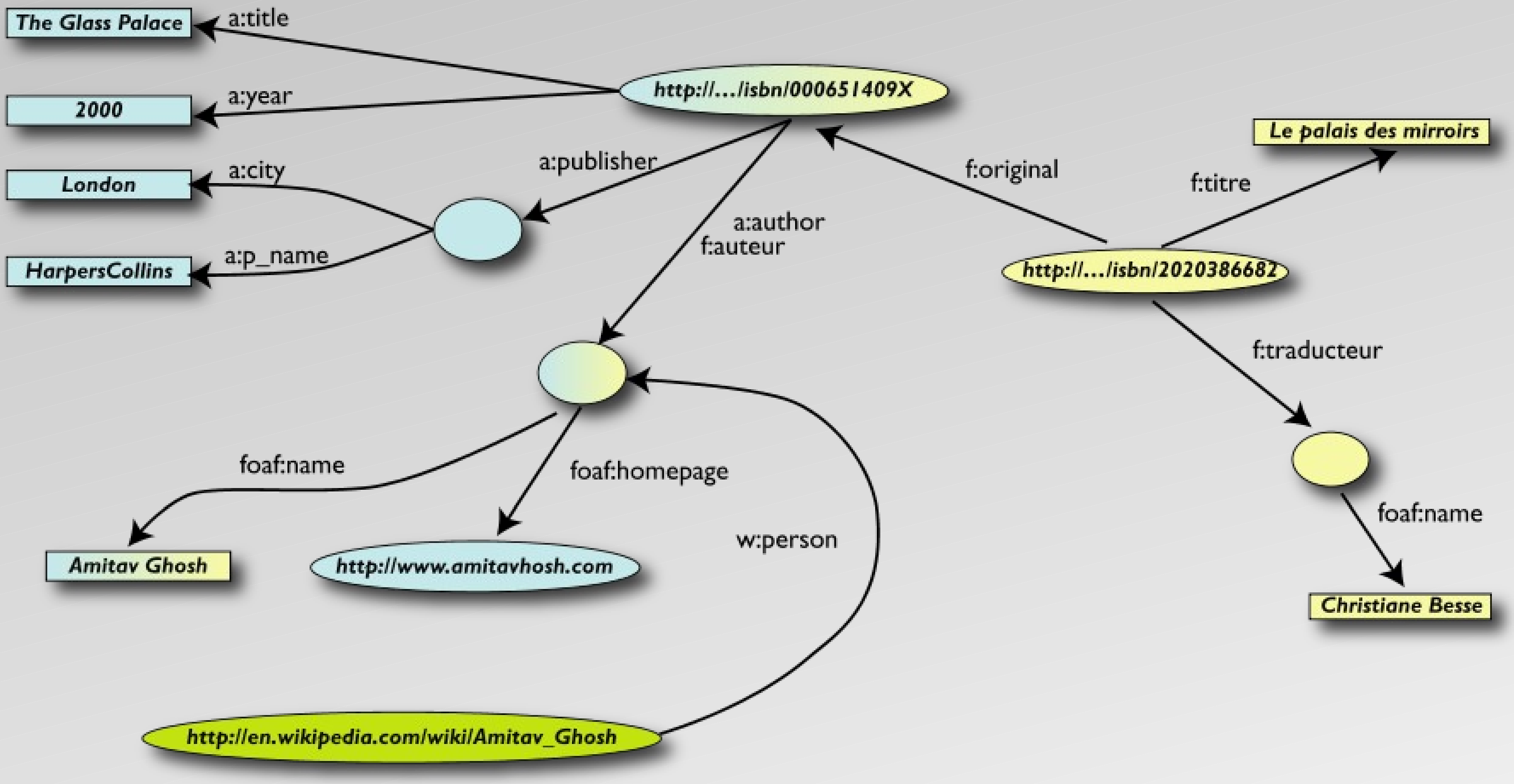
- User of dataset "F" can now query:
  - "give me the home page of the original's author"
- The information is not in datasets "F" or "A"…
- …but was made available by:
  - merging datasets "A" and datasets "F"
  - adding three simple extra statements as an extra "glue"



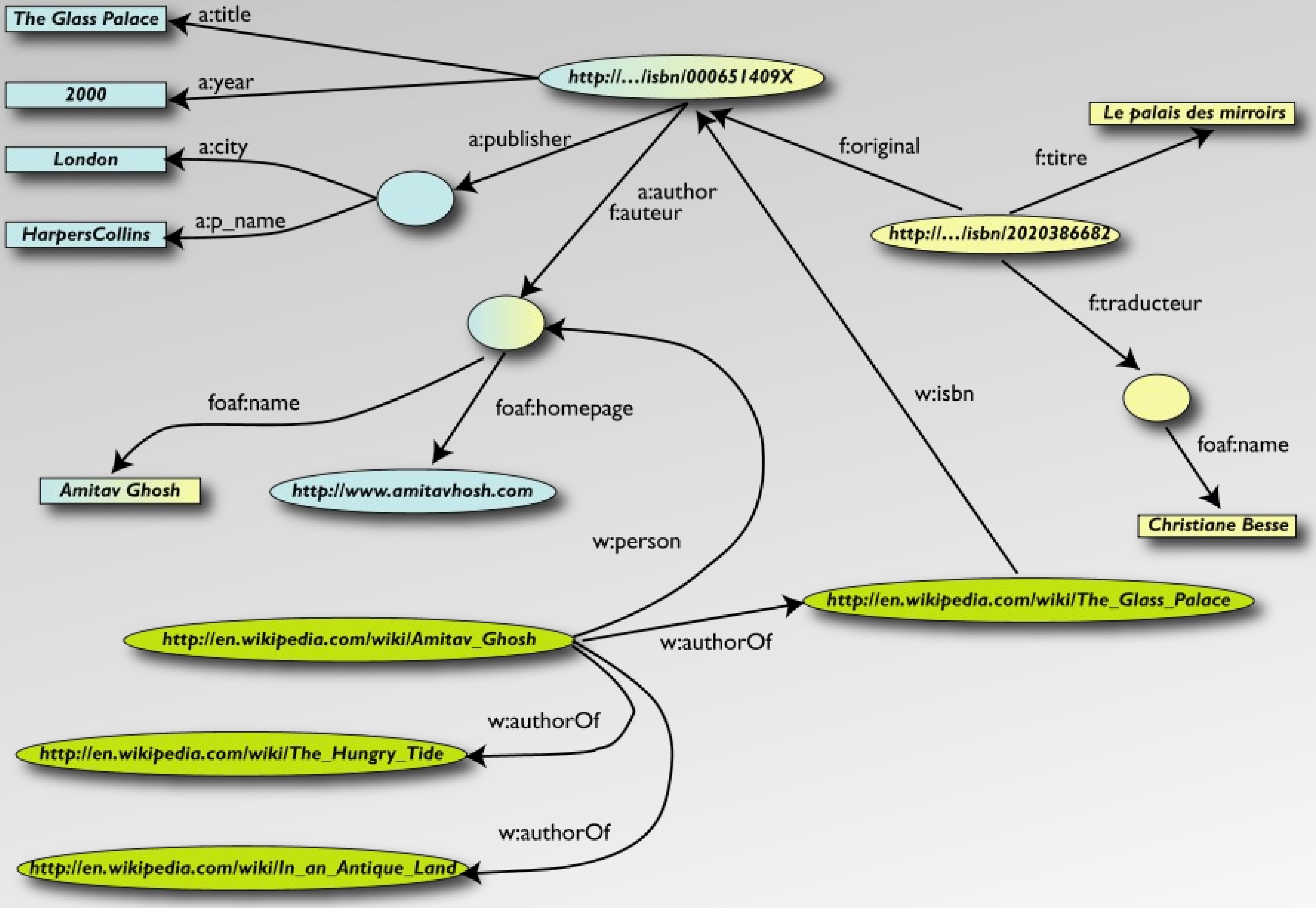
Combine with different datasets

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

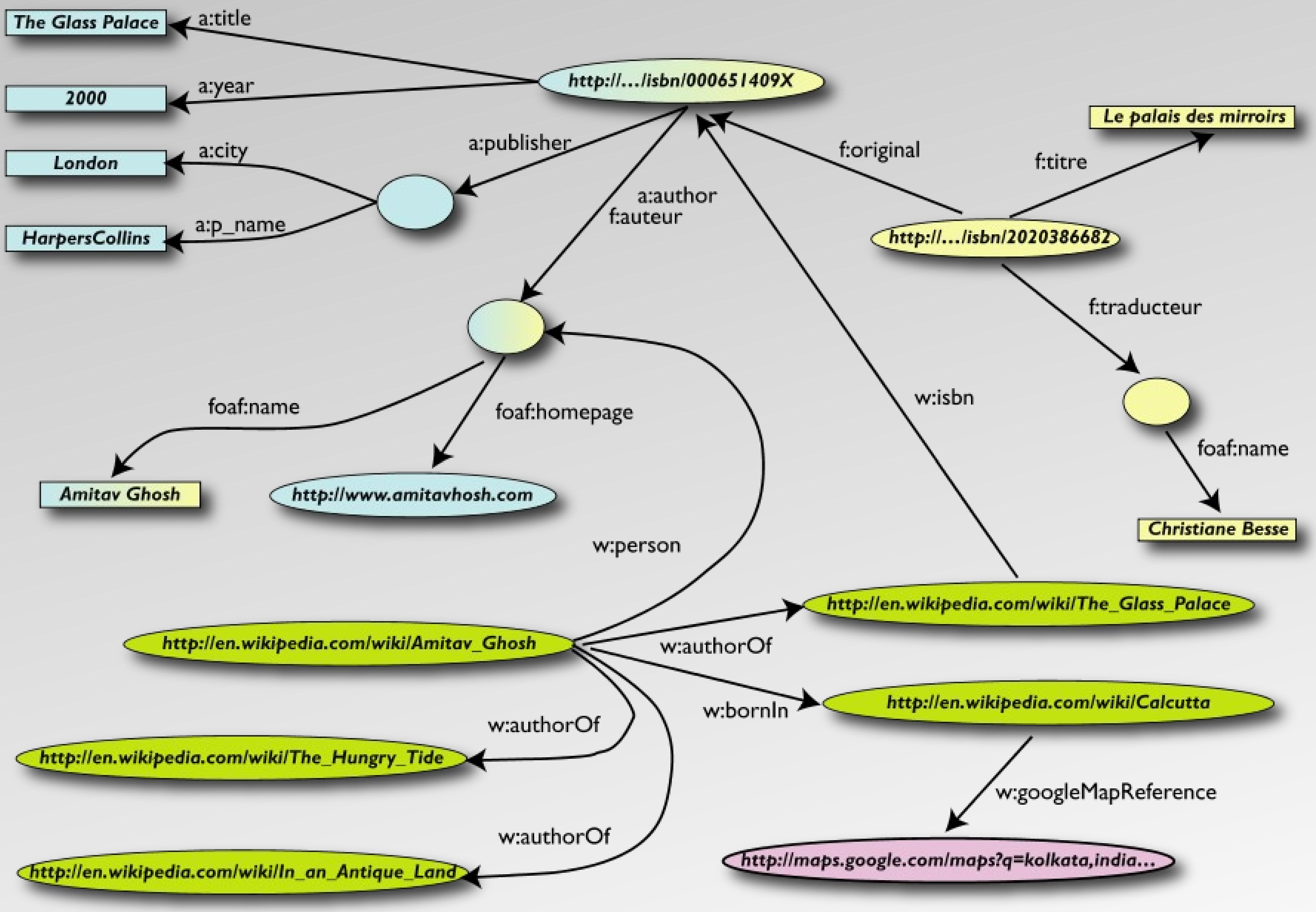
Merge with Wikipedia data



Merge with Wikipedia data



Merge with Wikipedia data


The Glass Palace
a:title
2000
a:year
http://.../isbn/000651409X
London
a:city
a:publisher
HarpersCollins
a:p_name
a:author
f:auteur
f:original
Le palais des mirroirs
f:titre
http://.../isbn/2020386682
f:traducteur
foaf:name
foaf:homepage
Amitav Ghosh
http://www.amitavhosh.com
w:isbn
foaf:name
Christiane Besse
w:person
http://en.wikipedia.com/wiki/The_Glass_Palace
http://en.wikipedia.com/wiki/Amitav_Ghosh
w:authorOf
http://en.wikipedia.com/wiki/Calcutta
w:bornIn
w:authorOf
http://en.wikipedia.com/wiki/The_Hungry_Tide
w:googleMapReference
w:authorOf
http://en.wikipedia.com/wiki/In_an_Antique_Land
http://maps.google.com/maps?q=kolkata,india...

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

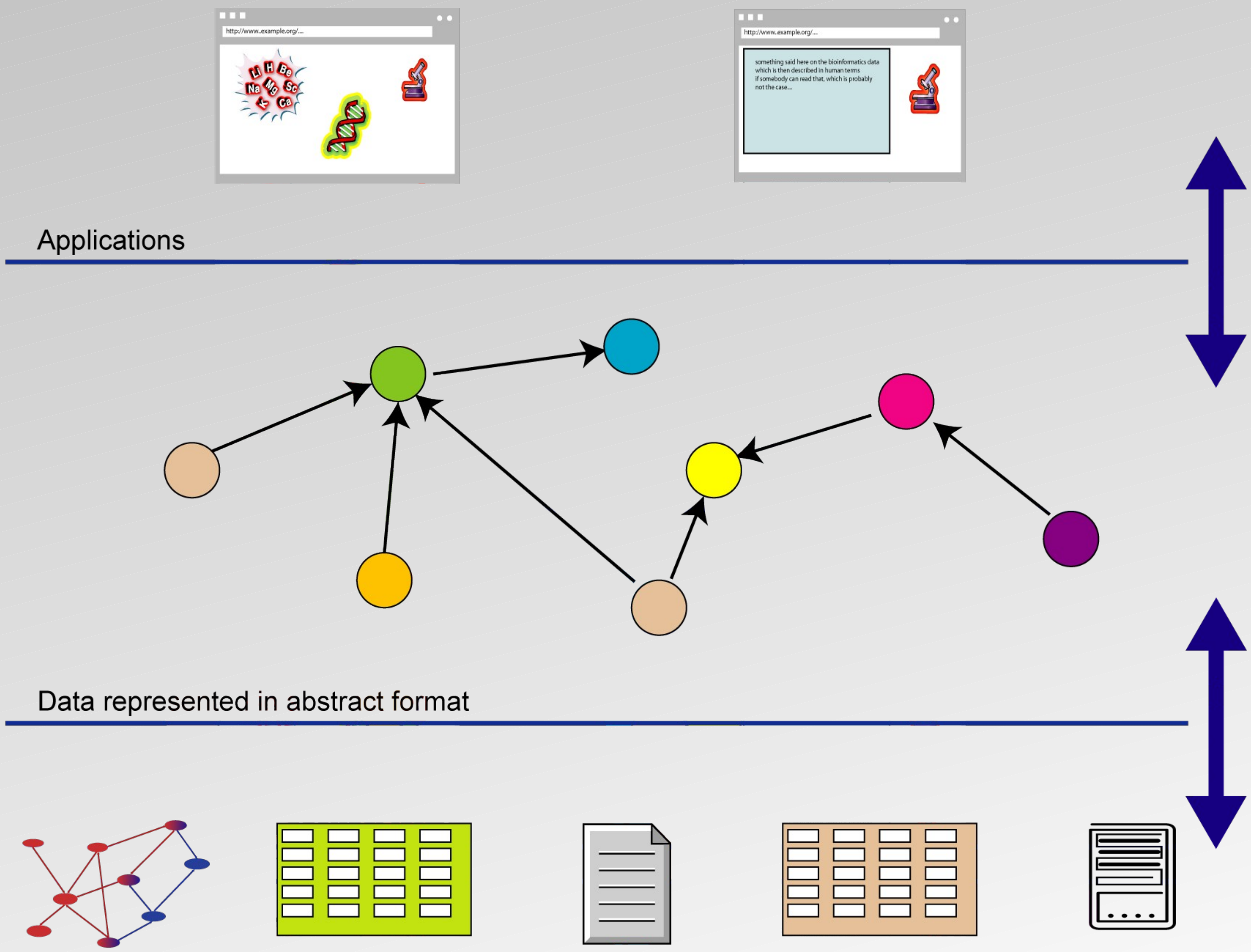
What did we do?

- 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 *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)


http://www..example.org/...
Li H Be Na Mg Sc K Ca
http://www..example.org/...
something said here on the bioinformatics data
which is then described in human terms
if somebody can read that, which is probably
not the case...

Applications

Query,
Manipulate,
etc.

Data represented in abstract format

Map,
Expose,
etc.

Data in various formats

The abstraction pays off because…

- … the graph representation is independent on the exact format, data structures, schemas
- … 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

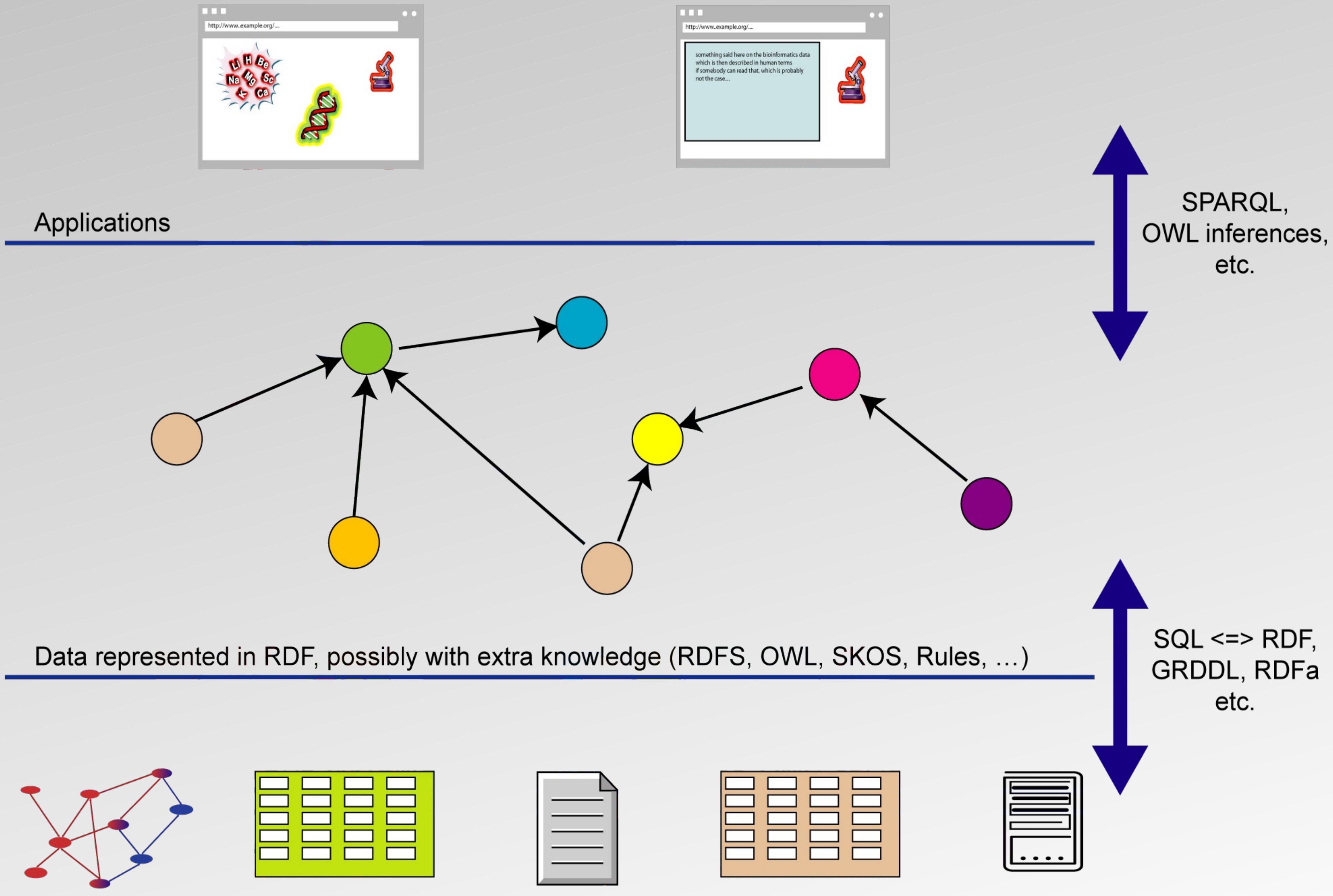
The network effect

- 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

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
  - extract RDF information from XML (eg, XHTML) pages: GRDDL
  - add structured information to XHTML pages: RDFa
  - a query language adapted for the relational graphs: SPARQL
  - 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)



something said here on the bioinformatics data which is then described in human terms if somebody can read that, which is probably not the case...

Applications

SPARQL, OWL inferences, etc.

Data represented in RDF, possibly with extra knowledge (RDFS, OWL, SKOS, Rules, ...)

SQL <=> RDF, GRDDL, RDFa etc.

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 2008”!

Semantic Web applications

- The data integration is only one area of SW applications
- Let us see some more…

Practical applications

- Follow the separate slide set

Conclusions

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

CEO guide for SW: the “DO-s”

- **Start small**: Test the Semantic Web waters with a pilot project […] before investing large sums of time and money.
- **Check credentials**: A lot of systems integrators don't really have the skills to deal with Semantic Web technologies. Get someone who's savvy in semantics.
- **Expect training challenges**: It often takes people a while to understand the technology. […]
- **Find an ally**: It can be hard to articulate the potential benefits, so find someone with a problem that can be solved with the Semantic Web and make that person a partner.

Source: BusinessWeek Online, April 2007

CEO guide for SW: the “DON’T-s”

- **Go it alone**: The Semantic Web is complex, and it's best to get help. […]
- **Forget privacy**: Just because you can gather and correlate data about employees doesn’t mean you should. Set usage guidelines to safeguard employee privacy.
- **Expect perfection**: While these technologies will help you find and correlate information more quickly, they’re far from perfect. Nothing can help if data are unreliable in the first place.
- **Be impatient**: One early adopter at NASA says that the potential benefits can justify the investments in time, money, and resources, but there must be a multi-year commitment to have any hope of success

Source: BusinessWeek Online, April 2007

Thank you for your attention!

- These slides are publicly available on:

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