What is the Semantic Web?

17th XBRL International Conference
Eindhoven, the Netherlands
5st May, 2008

Ivan Herman, W3C

> Towards a Semantic Web

- The current Web represents information using
  - natural language (English, Hungarian, Dutch,…)
  - 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?
  - …

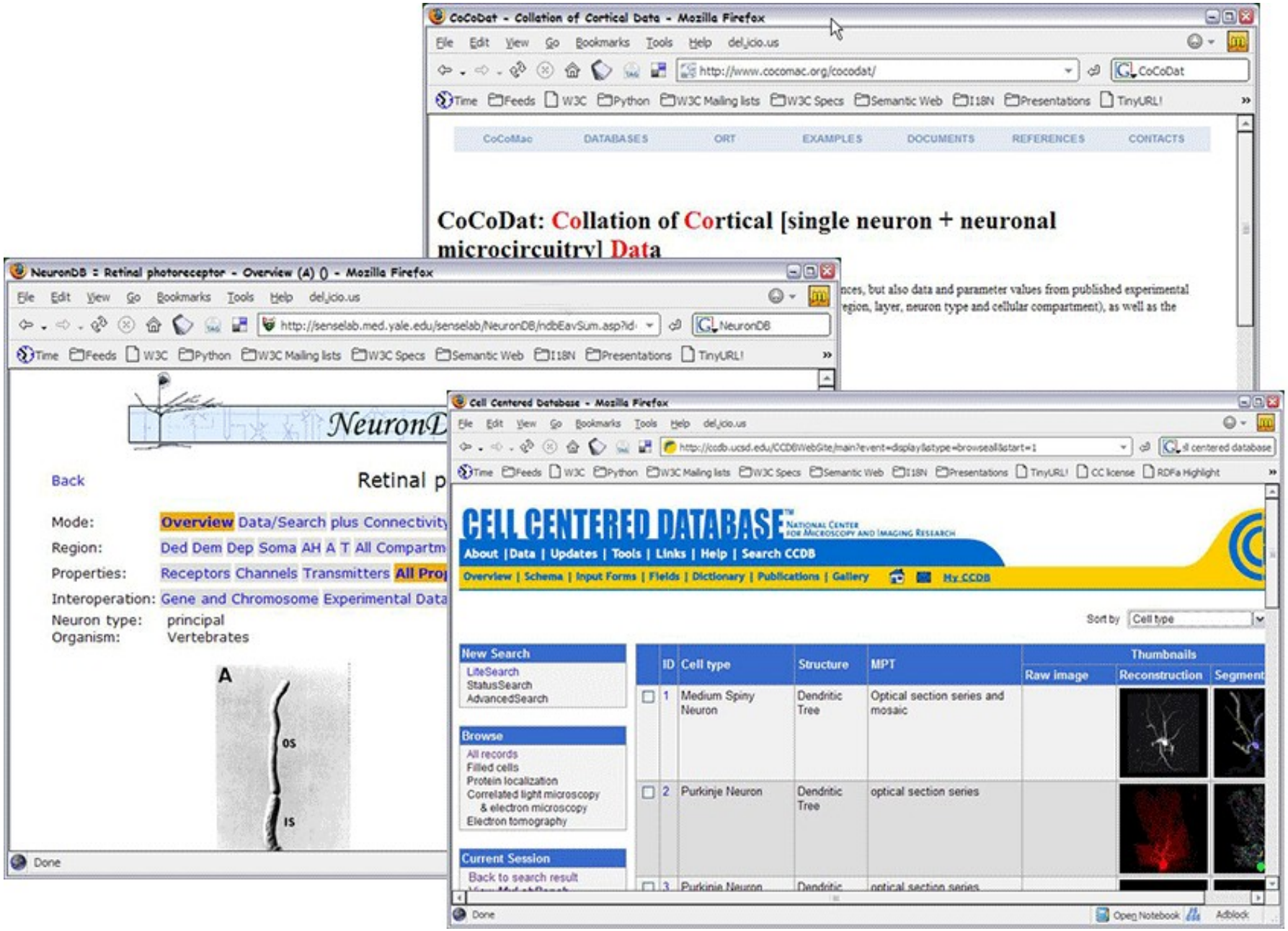
> Example: automatic airline reservation

- 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…



> 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

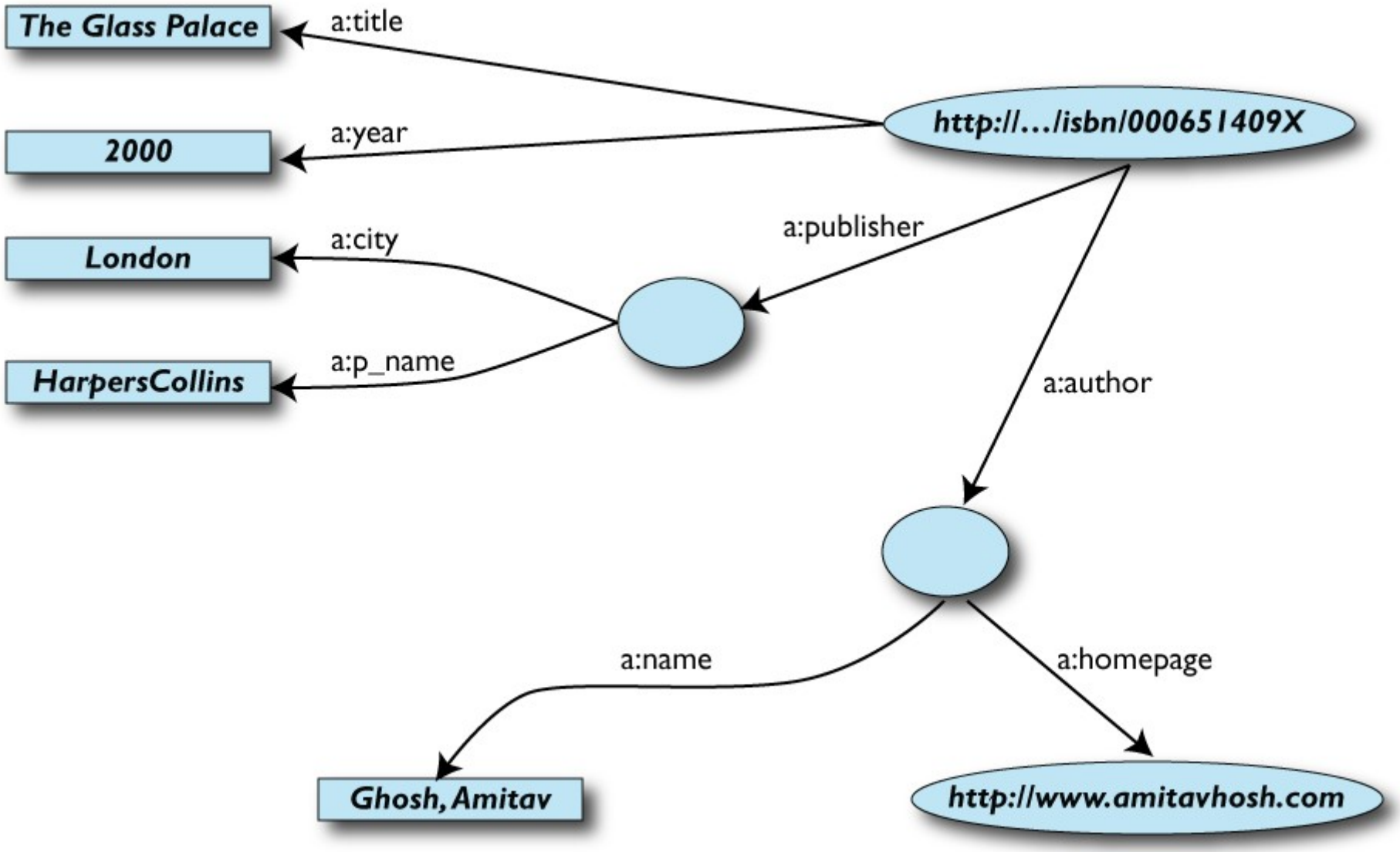
> A *simplified* bookstore data (dataset "A")

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

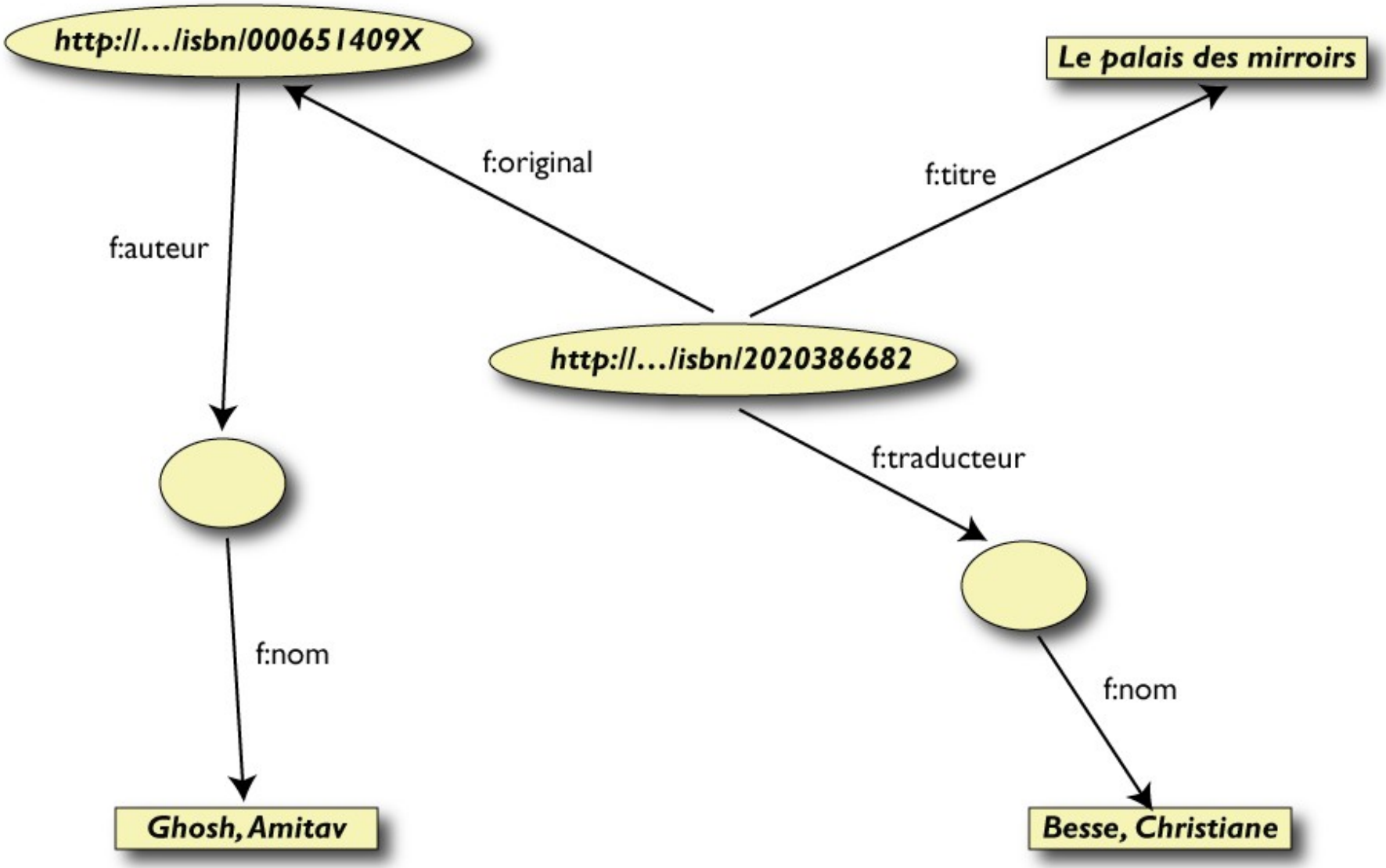
- Relations form a graph
  - the nodes refer to the “real” data or contain some literal
  - how the graph is represented in machine is immaterial for now
- Data export does *not* necessarily mean physical conversion of the data
  - relations can be generated on-the-fly at query time
    - via SQL “bridges”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

> Another bookstore data (dataset “F”)

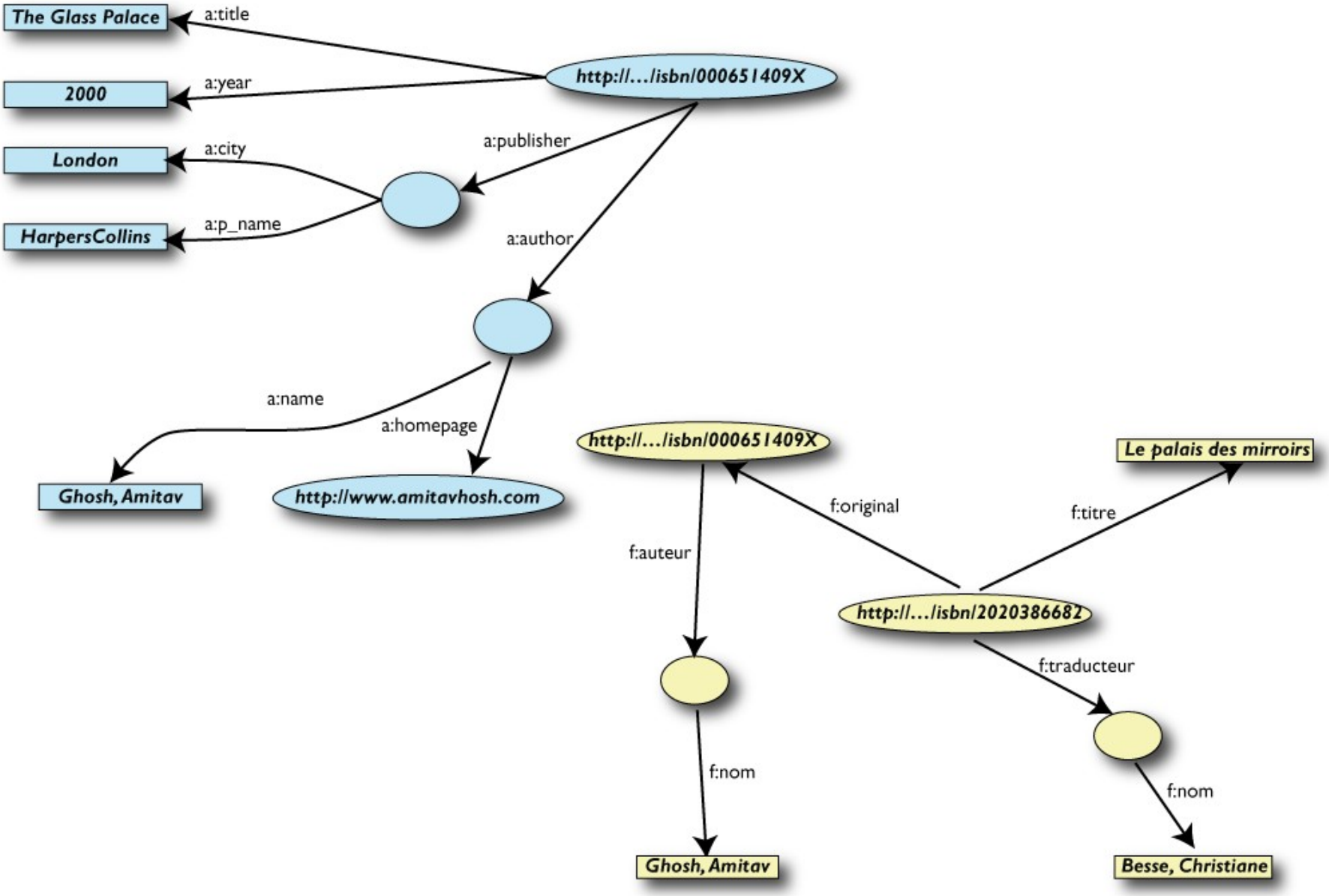
	A	B	C	D	E
1	**ID**	**Titre**	**Auteur**	**Traducteur**	**Original**
2	ISBN0 2020386682	Le Palais des miroirs	A7	A8	ISBN-0-00-651409-X
3					
4					
5					

	Nom
6	**Nom**
7	Ghosh, Amitav
8	Besse, Christianne

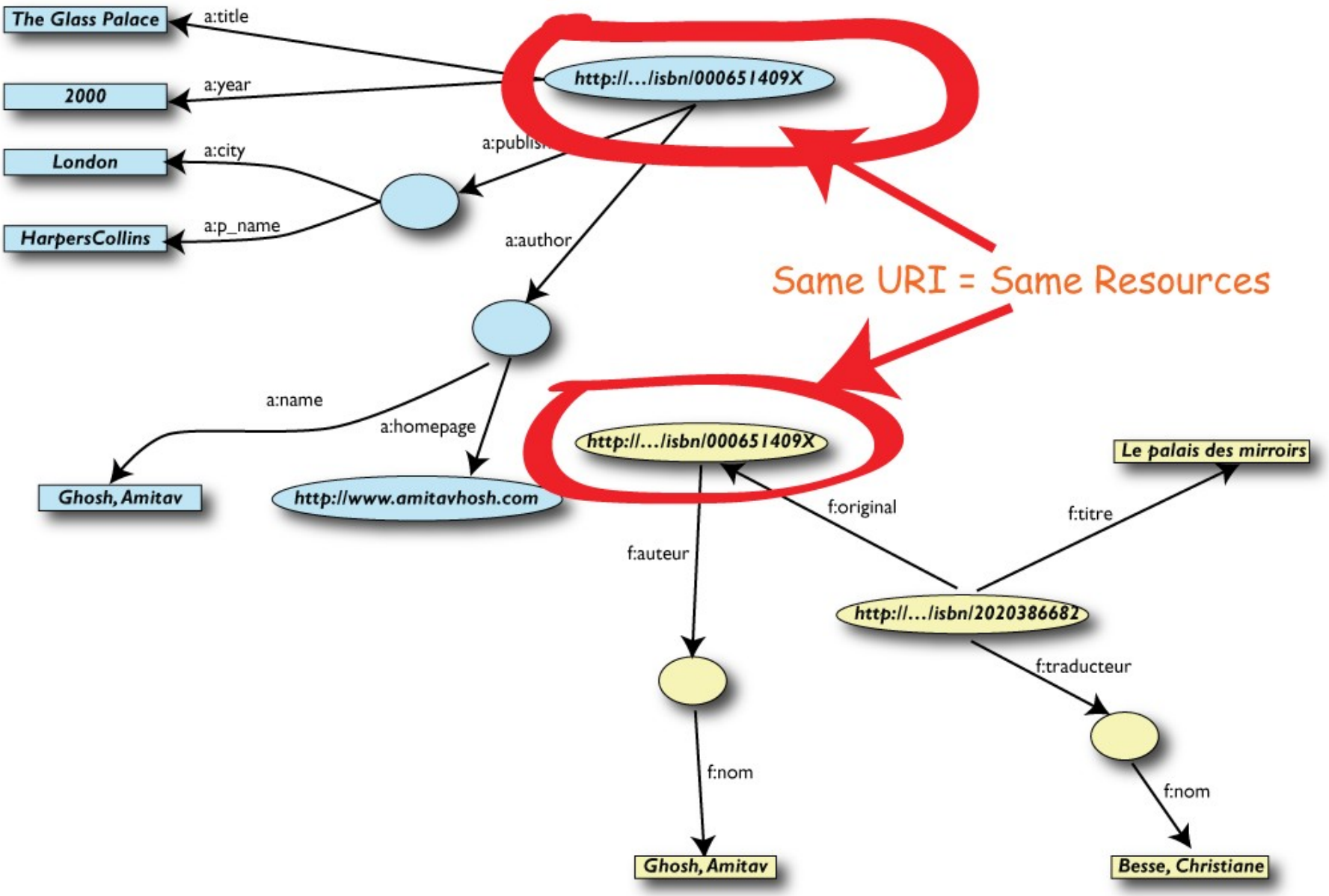
> 2nd: export your second set of data



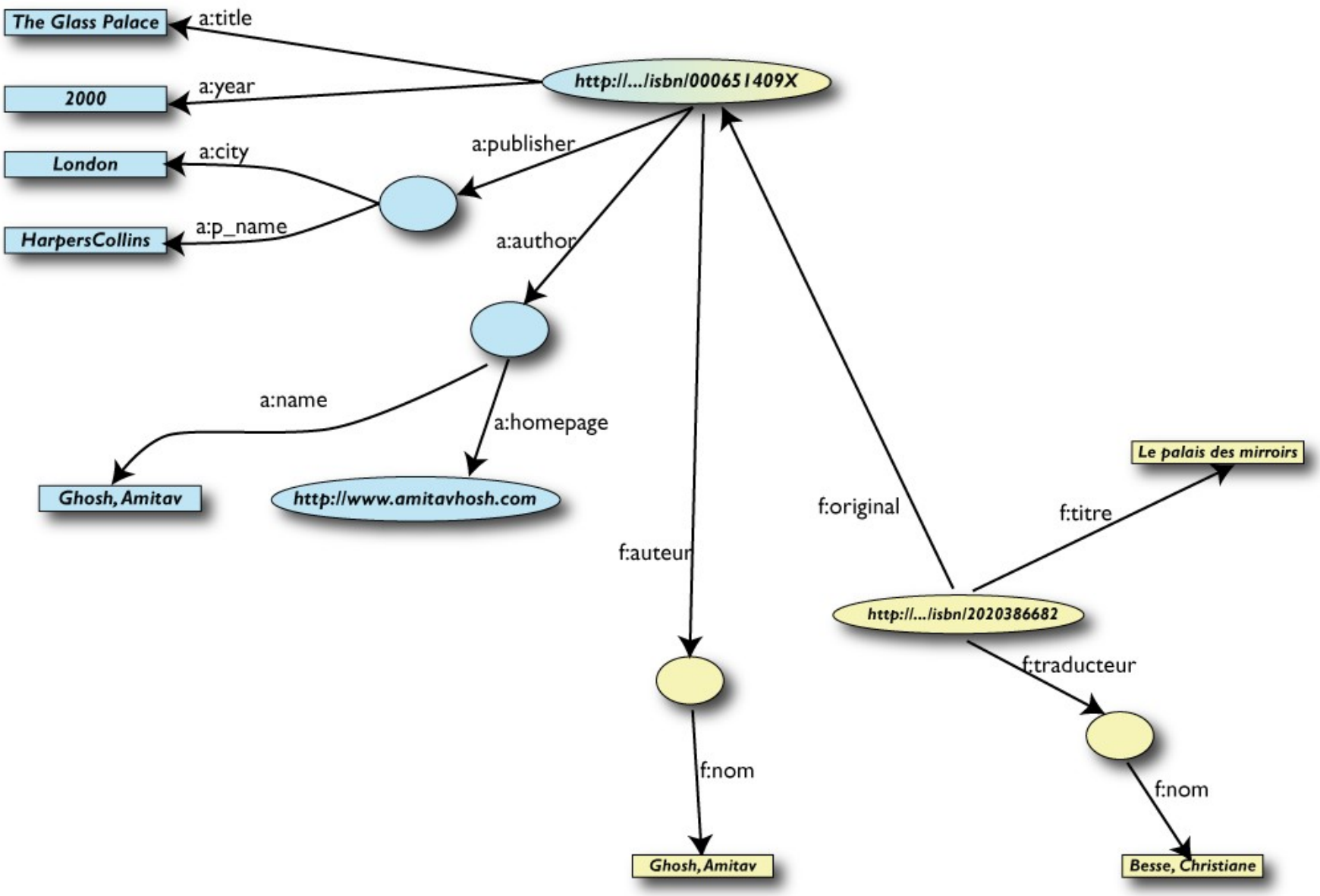
> 3rd: start merging your data



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

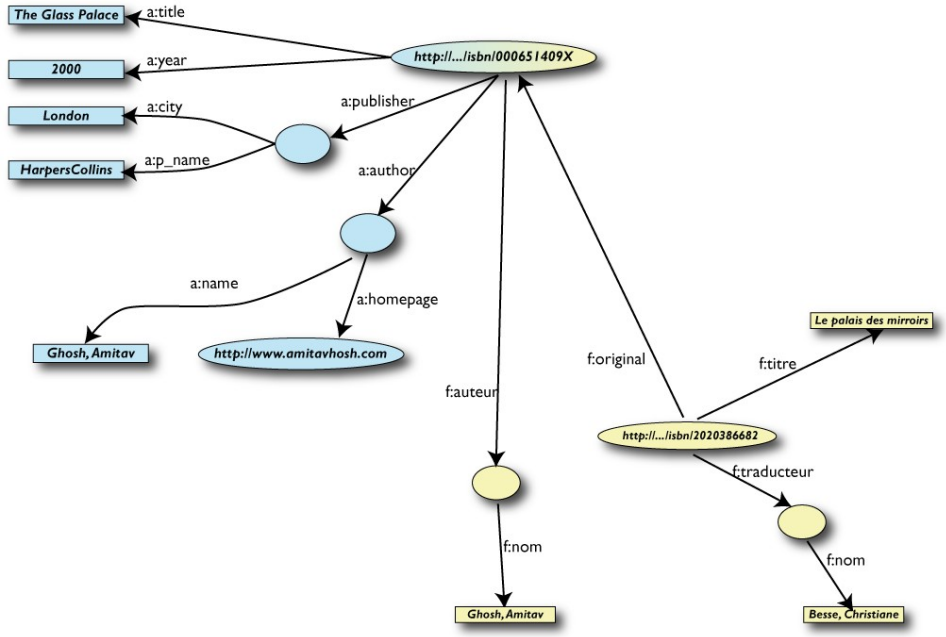


> 3rd: merge identical resources



> Start making queries…

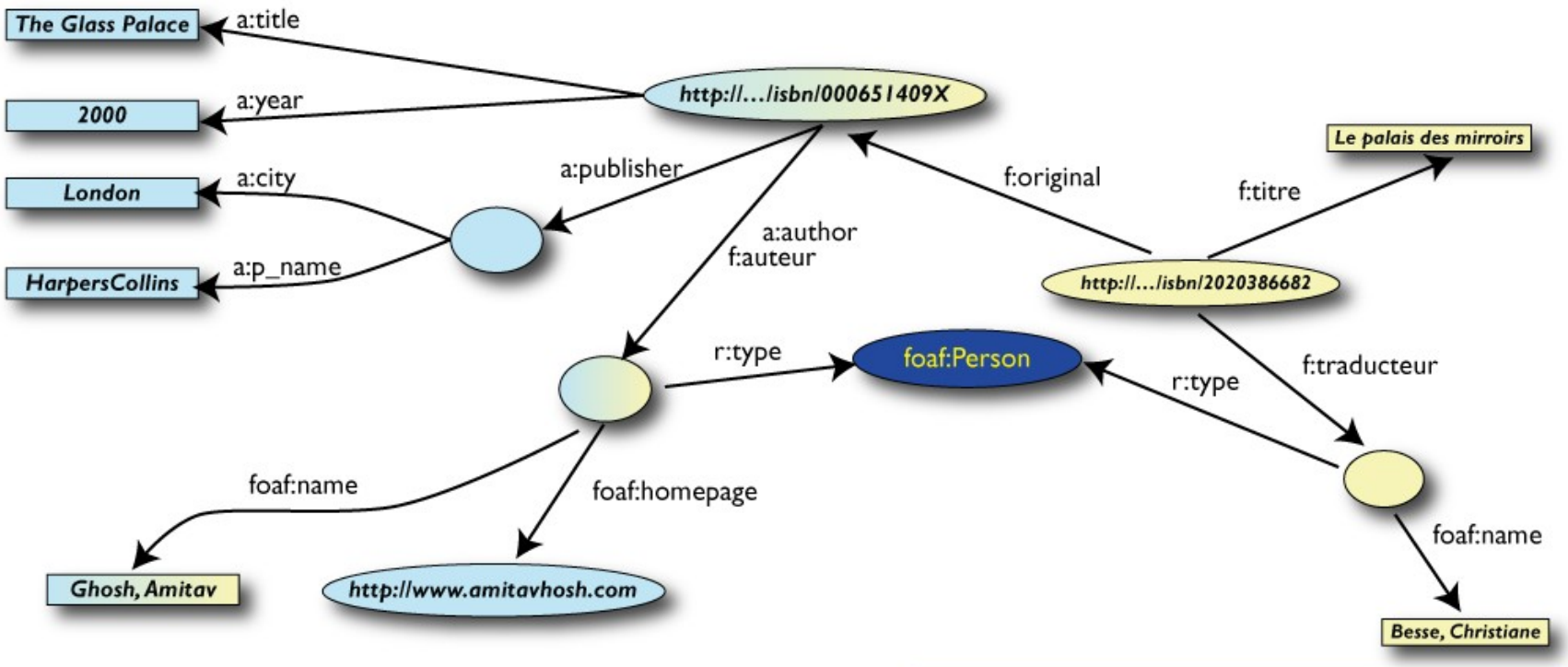
- User of data “F” can now ask queries like:
  - « donnes-moi le titre de l’original »
  - (ie: “give me the title of the original”)
- This information is not in the dataset “F”…
- …but can be retrieved by merging with dataset “A”!

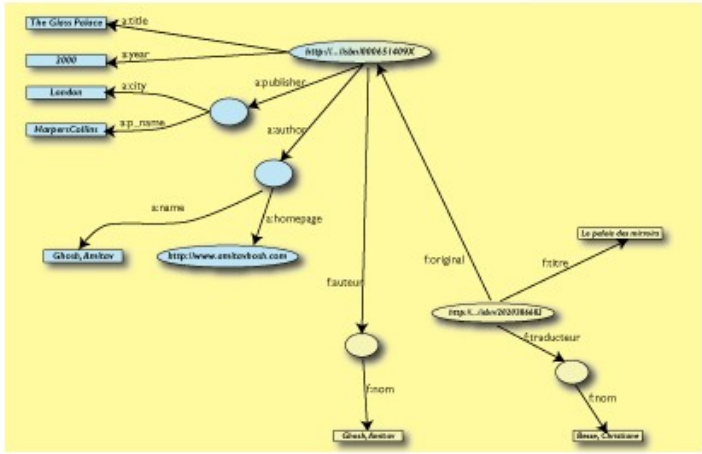


> However, more can be achieved…

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

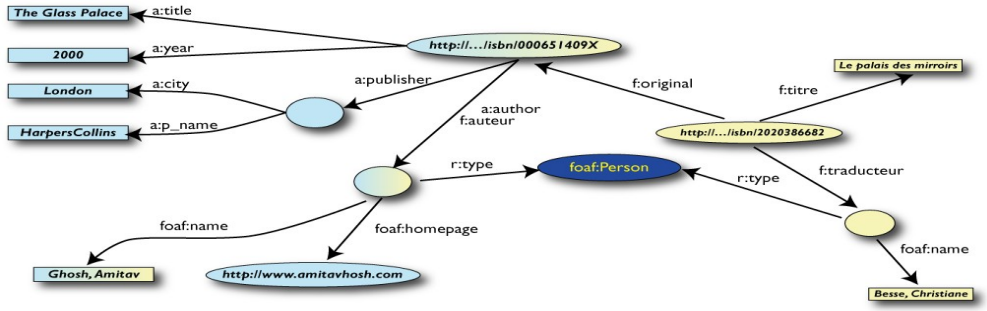
> 3rd revisited: use the extra knowledge





> Start making richer queries!

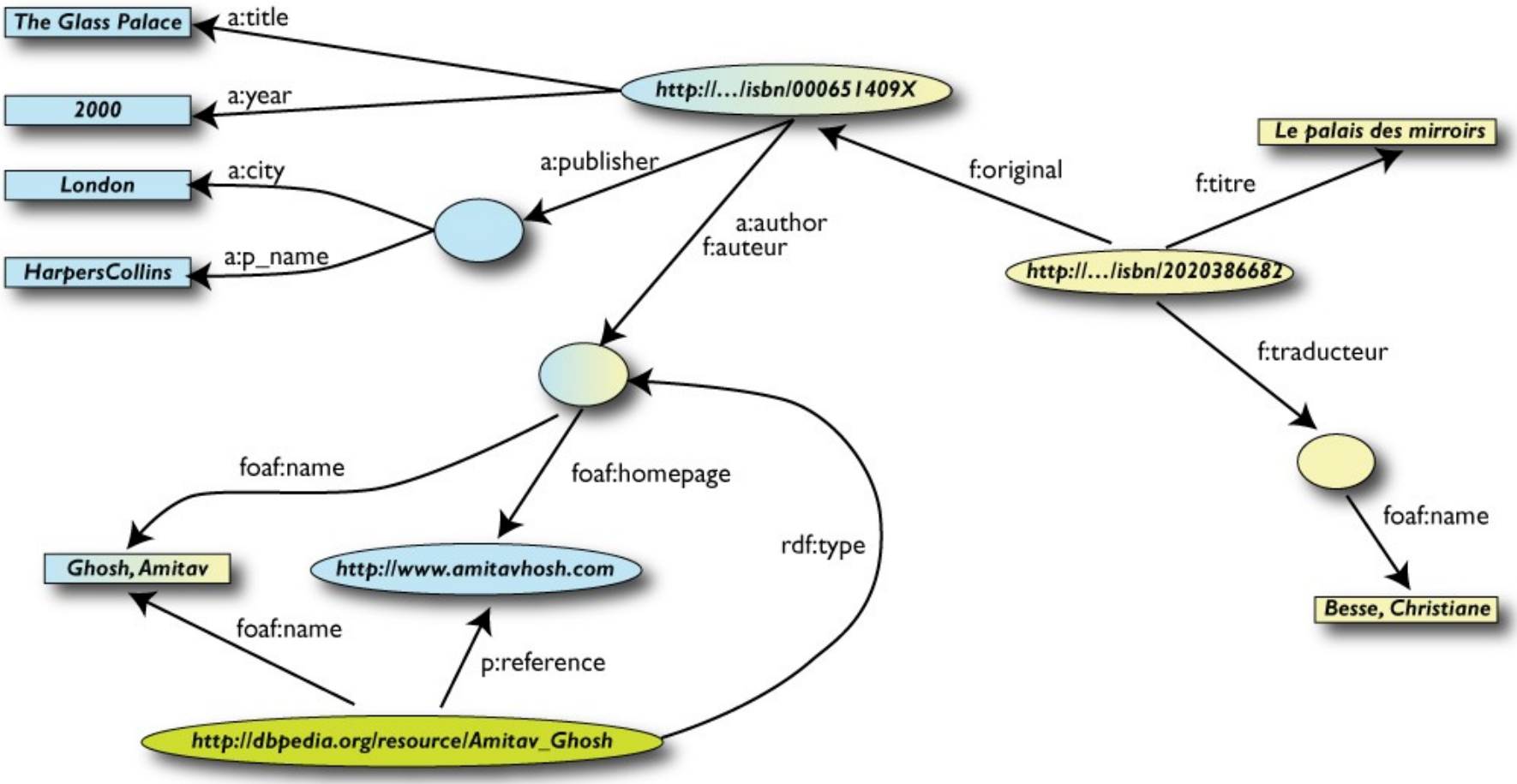
- User of dataset “F” can now query:
  - «donnes-moi la page d’accueil de l’auteur de l’original»
  - (ie, “give me the home page of the original’s author”)
- The information is not in datasets “F” or “A”…
- …but was made available by:
  - merging datasets “A” and datasets “F”
  - adding three simple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



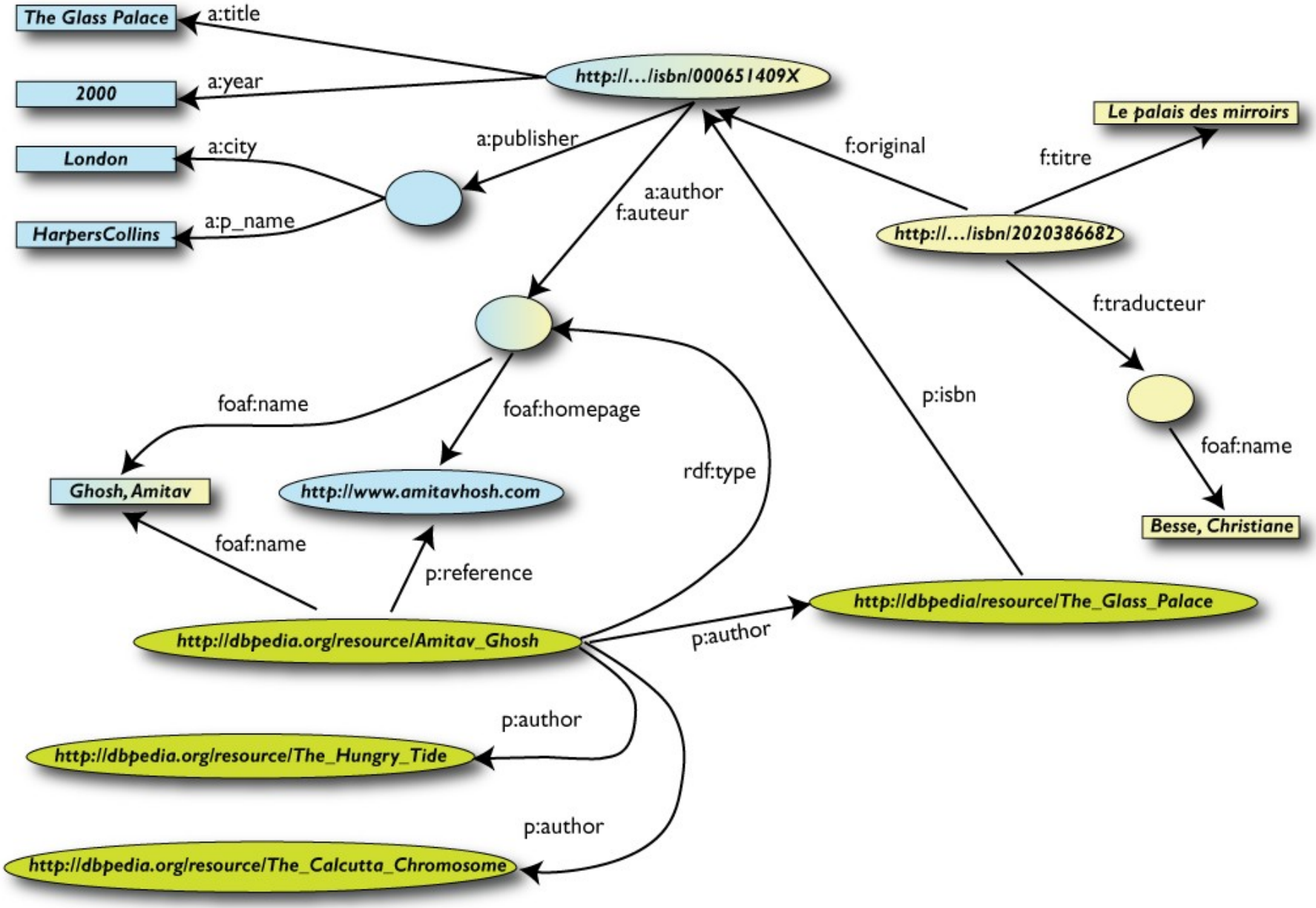
> Combine with different datasets

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

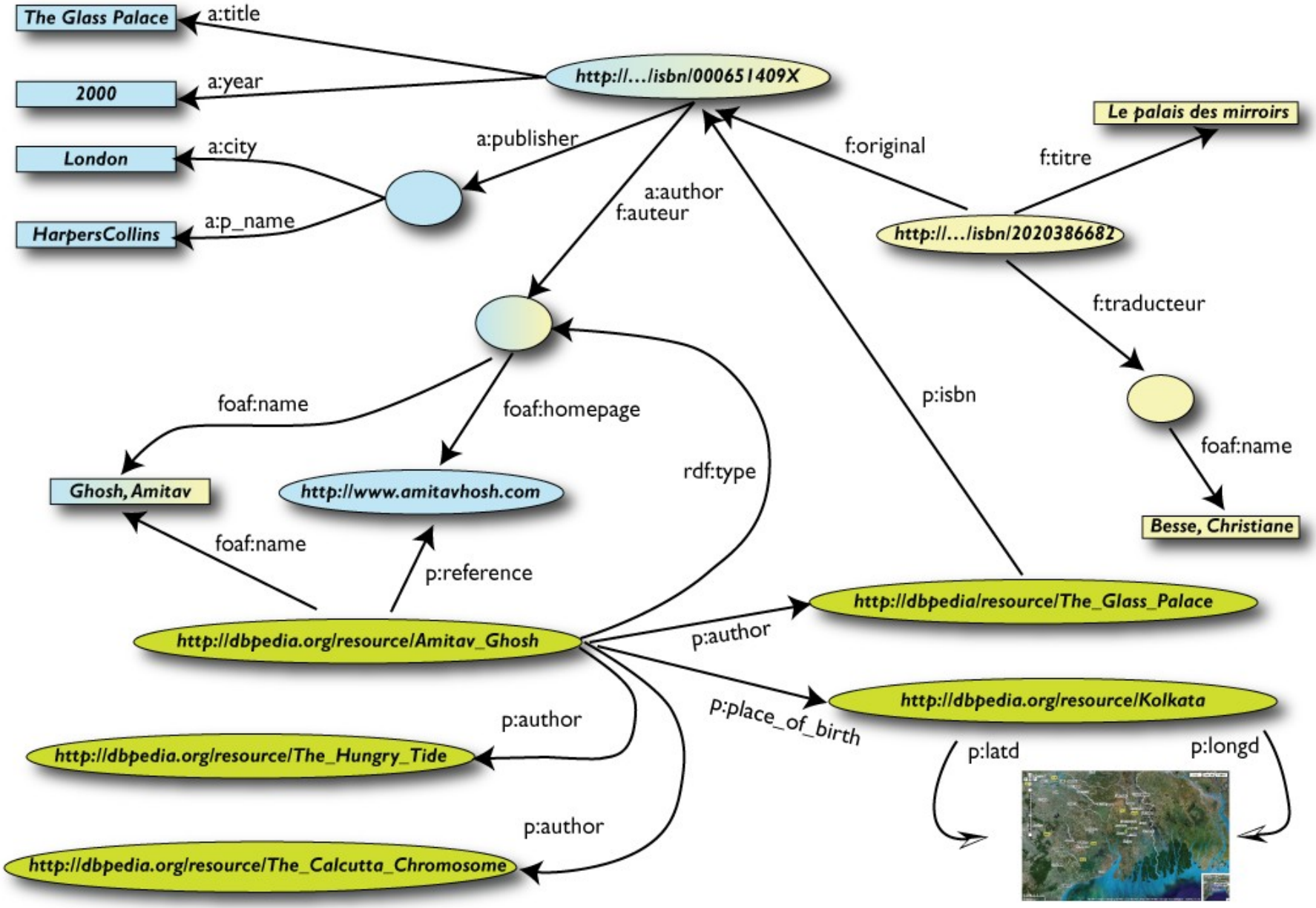
> Merge with Wikipedia data



> 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

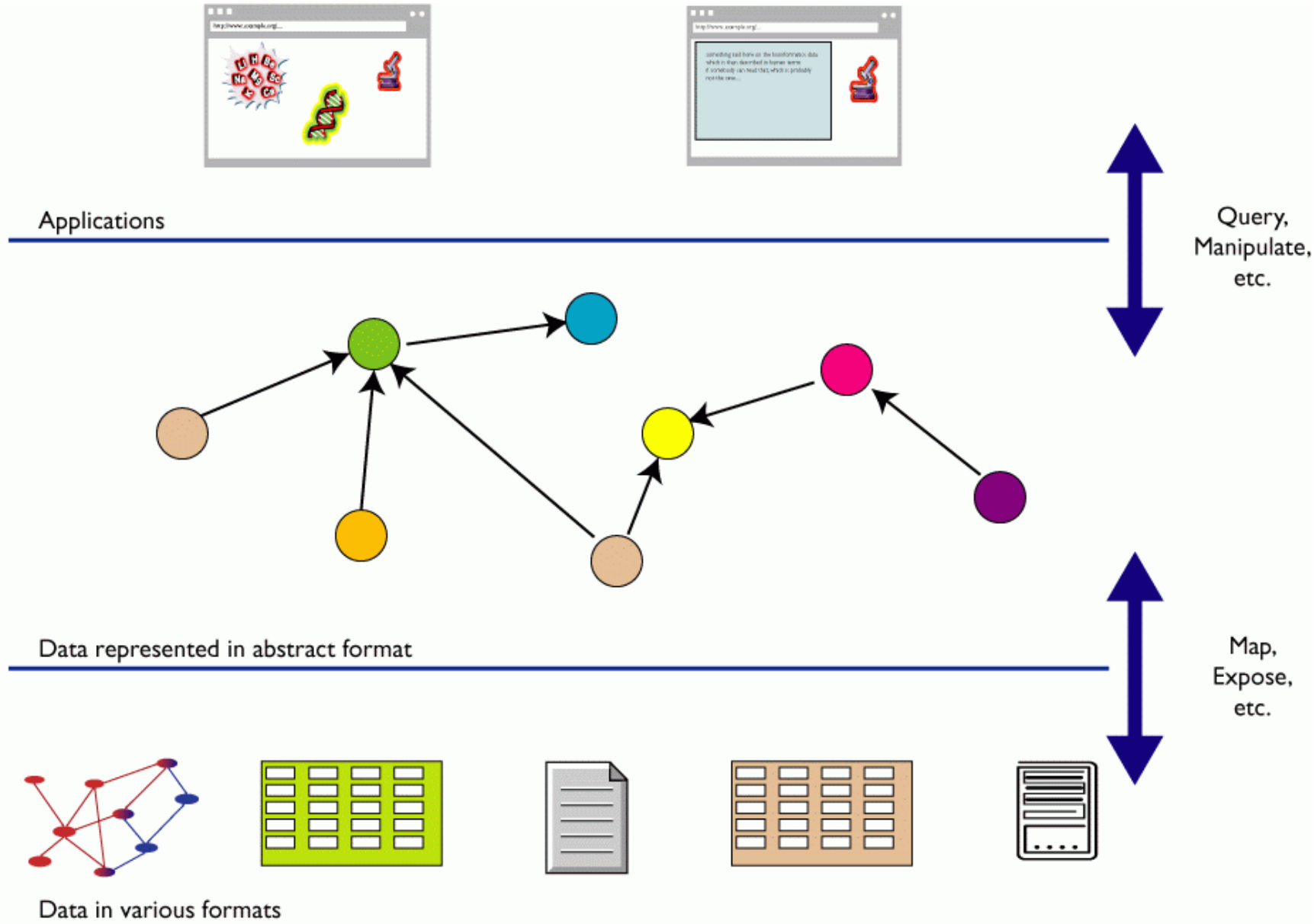
> What did we do?

- We combined different datasets 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)



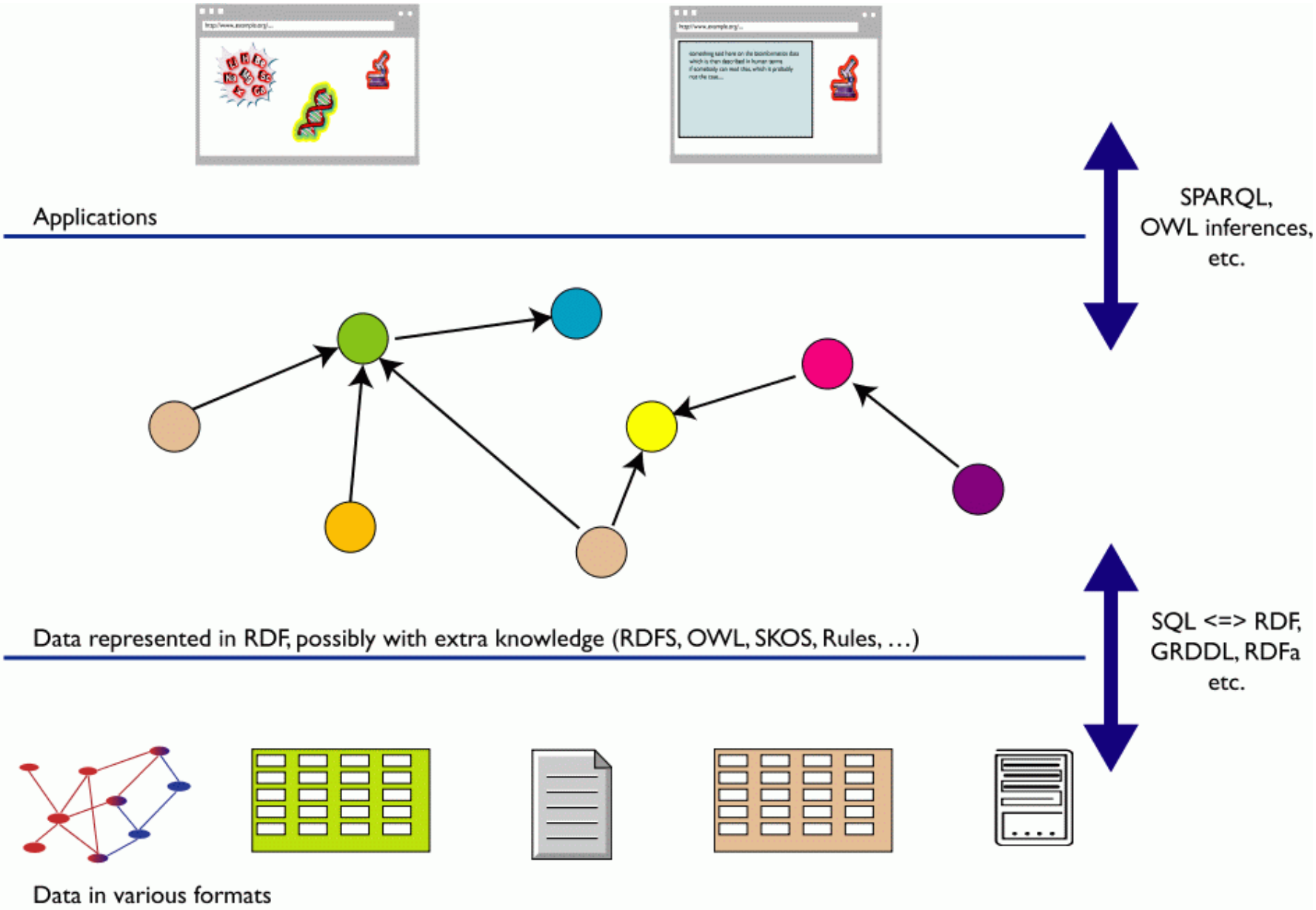
> The abstraction pays off because…

- … 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
- … new data, new connections can be added seamlessly, regardless of the structure of other data sources

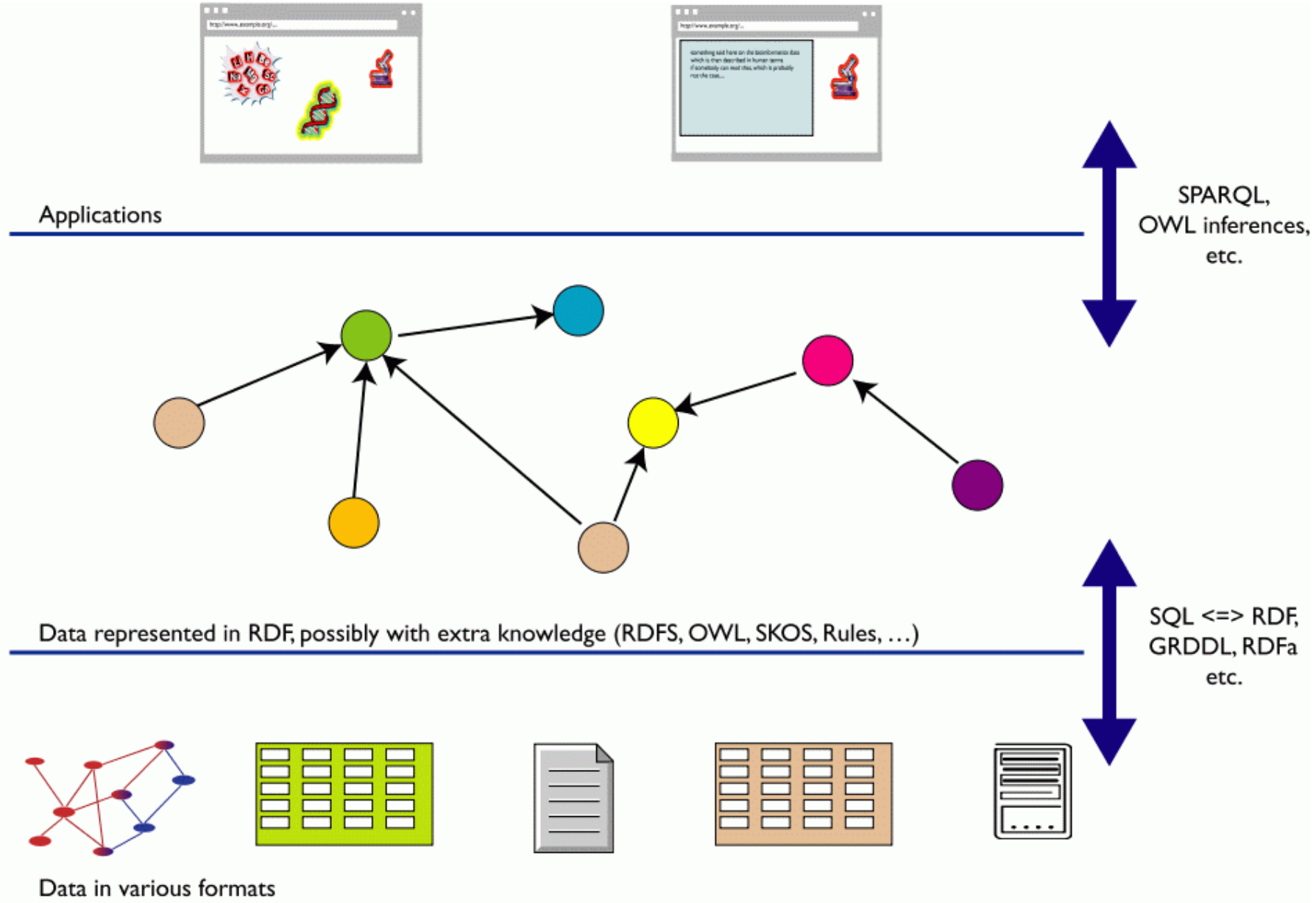
> So where is the Semantic Web?

- The Semantic Web provides technologies to make such integration possible! For example:
  - an abstract model for the relational graphs: RDF (with different “serializations” in XML or text)
  - extract RDF information from XML data: GRDDL
  - 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)

> How they fit on the picture…



> So where is the Semantic Web? (cont)



> Public datasets are accumulating

- "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
- These data are not only available for the Semantic Web, but they are also fully public…

> And XBRL?

- An outsider's view, of course…
- The XBRL spec achieves a major integration of data… but only within a specific domain
- If the financial data is to be combined with, say, statistical data: "bridging" XBRL to the Semantic Web might be a good approach
- It is not easy (XBRL seems fairly complex) but it might be worth it!

> Thank you for your attention!

- These slides are publicly available on:

http://www.w3.org/2008/Talks/0505-Eindhoven-IH/