BASI DI DATI II − 2 modulo
Parte X: Introduzione al Web Semantico

Based on a tutorial of Ivan Herman, W3C

Prof. Riccardo Torlone

Università Roma Tre

Outline

- Introduction
- The Web of Data
- Semantic Web
- Technologies

Using the Web

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

- First: 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

Using the Web

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

- First: you try to find a proper flight with …
- Second: you have to find a hotel, so you look for…

… a really cheap accommodation, or …

… or a really luxurious one, or …

… an intermediate one …

Using the Web

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

- First: you try to find a proper flight with …
- Second: you have to find a hotel, so you look for…
  - oops, that is no good, the page is in Hungarian that almost nobody understands, but…

… this one could work

Using the Web

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

- First: you try to find a proper flight with …
- Second: you have to find a hotel, so you look for…
  - Of course, you could decide to trust a specialized site…

… like this one, or…

… or this one

Using the Web

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

- First: you try to find a proper flight with …
- Second: you have to find a hotel, so you look for…
- Third: you may want to know something about Budapest; look for some photographs…

… on flickr …

… on Google …

… or 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!

Hidden Web

- 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

Some solution exists

- 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…

Another example

Social sites

- I have a list of "friends" by…

… Plaxo,

… Linkedin,

… Twitter,

...Xing,

… and, of course, Facebook

Problems here too

- 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?

Mashup solution

- 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

My home page

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, 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
  - □ make it possible to "link" to that URI from other sources of data (not only Web pages)
    - i.e., 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…

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

Human interpretation

- A human understands that this is my institution's home page

- He/she knows what it means (realizes that it is a University Department in Rome)

- 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

- It is that simple….the devil is in the details

  - a common model has to be provided for machines to describe, query, etc, the data and their connections

  - the "classification" of the terms can become very complex for specific knowledge areas: this is where ontologies, thesauri, etc, enter the game…

The rough structure of data integration

- Map the various data onto an abstract data representation
  - make the data independent of its internal representation…
- Merge the resulting representations
- 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
ISBN-00-0651049-X	id_auth1	The glass Palace	id_pub1	2000

ID	AUTHOR NAME	HOMEPAGE
id_auth1	Amitav Gosh	http://www.goshamitav.com

ID	PUBLISHER NAME	CITY
id_pub1	Harper Collins	London

1st: export your data as a set of relations
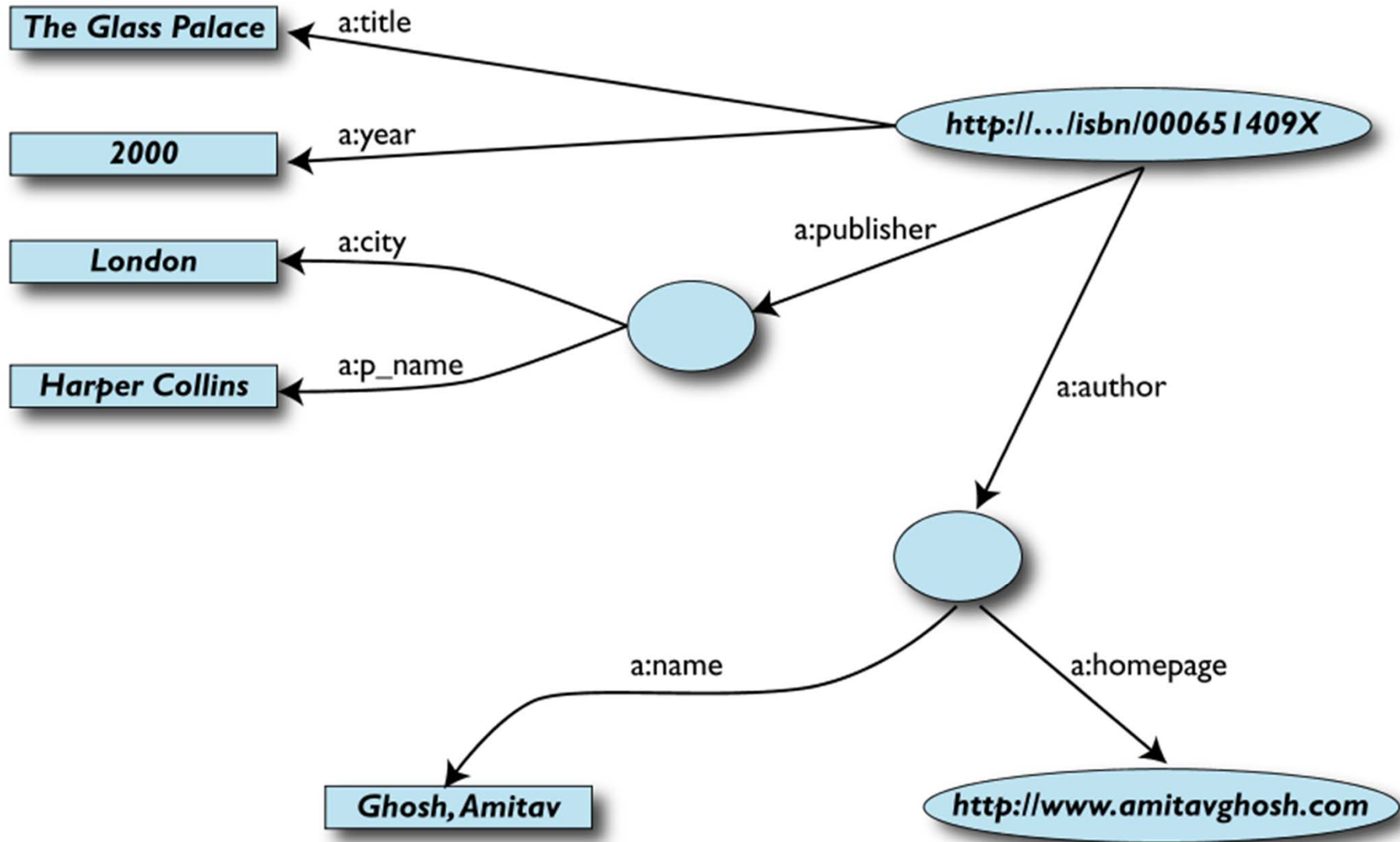
Some notes on the exporting the data

- **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")

ID	TITLE	TRADUCTER	ORIGINAL
ISBN-02-05668976-Y	The palais des miroirs	A13	ISBN-00-0651049-X

ID	AUTEUR
ISBN-00-0651049-X	A12

ID	NOM
A12	Amitav Gosh
A13	Christianne Bess

2nd: export your second set of data
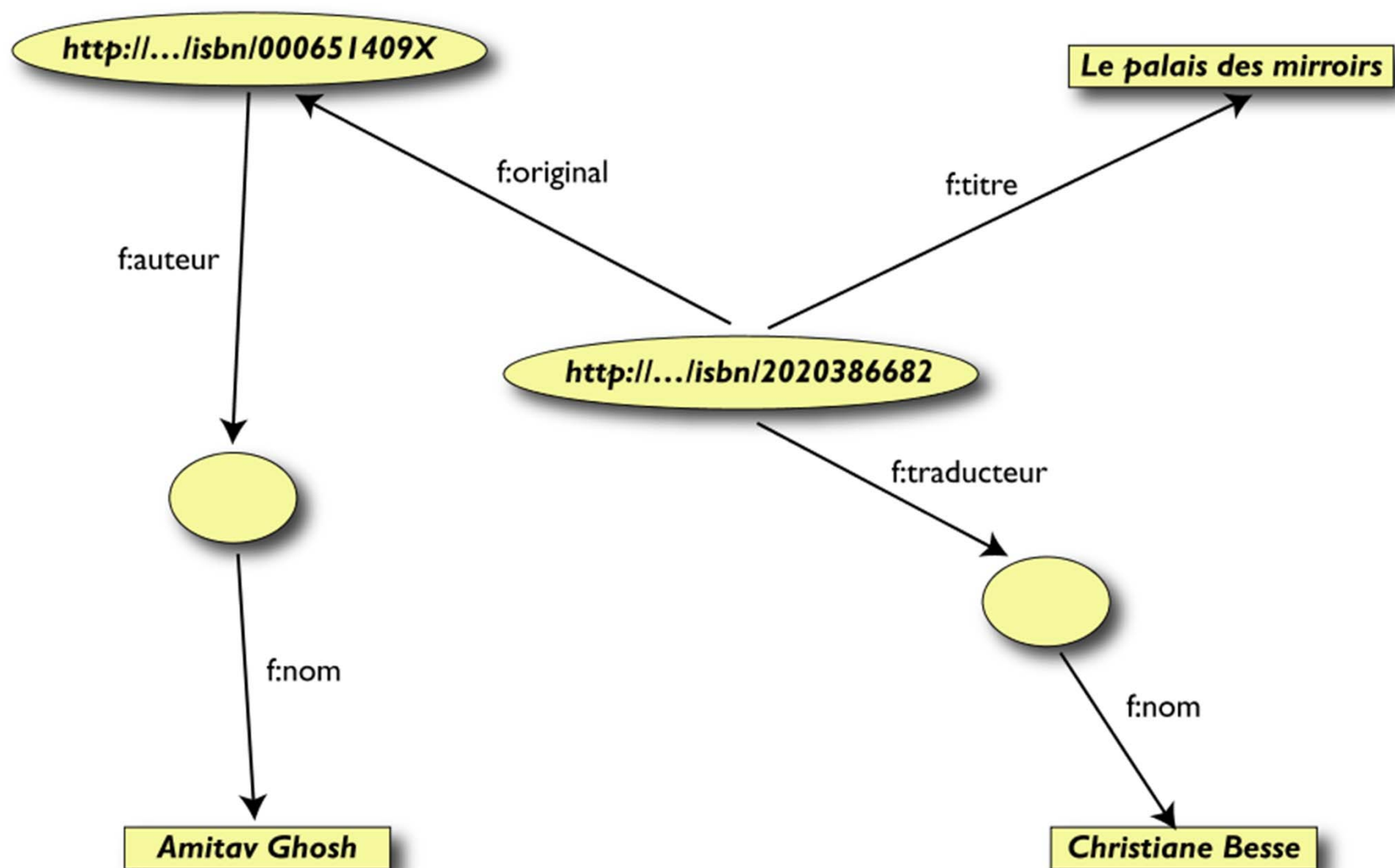
3rd: start merging your data
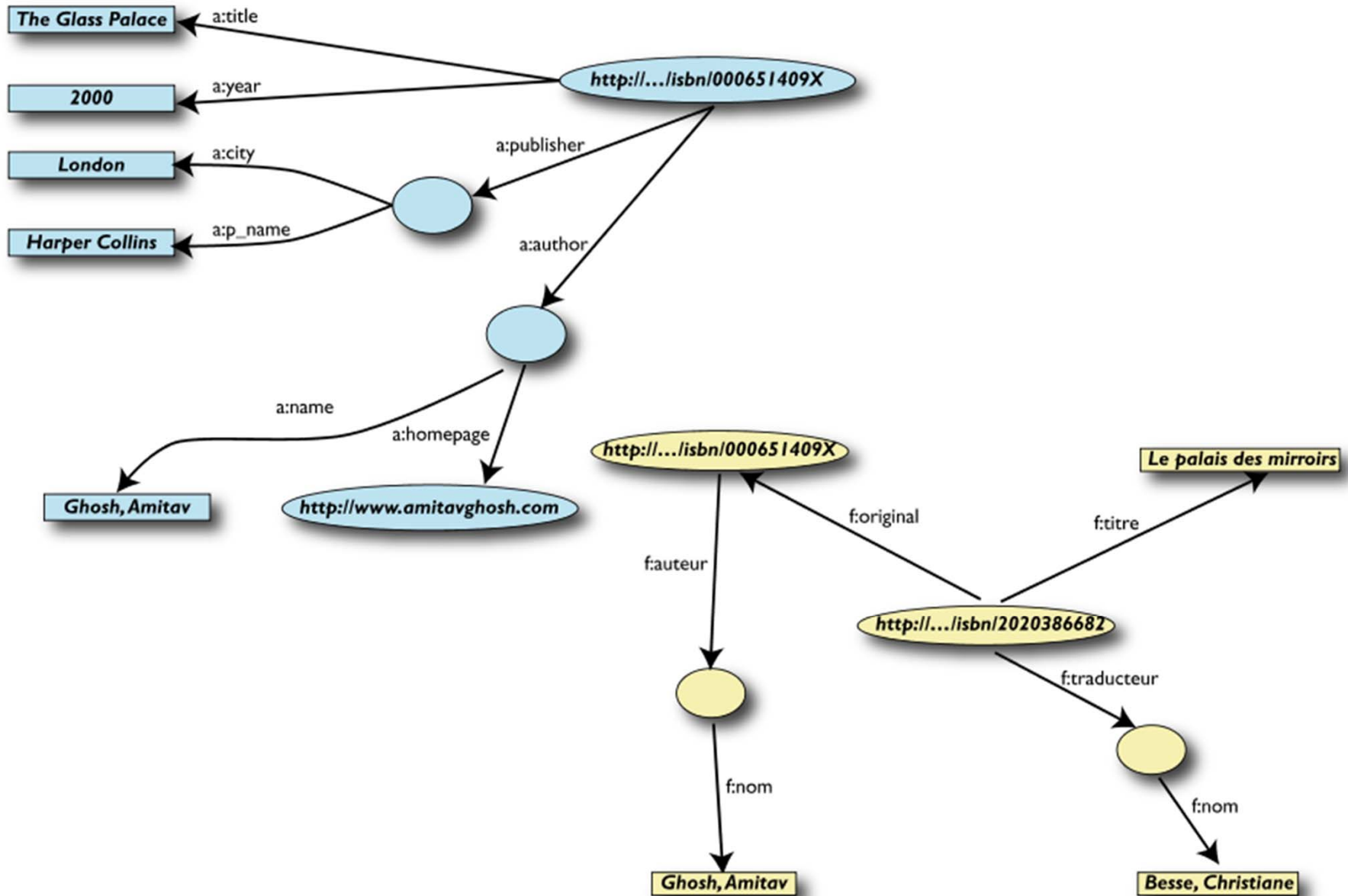
3rd: start merging your data (cont.)

Same URI = Same Resources
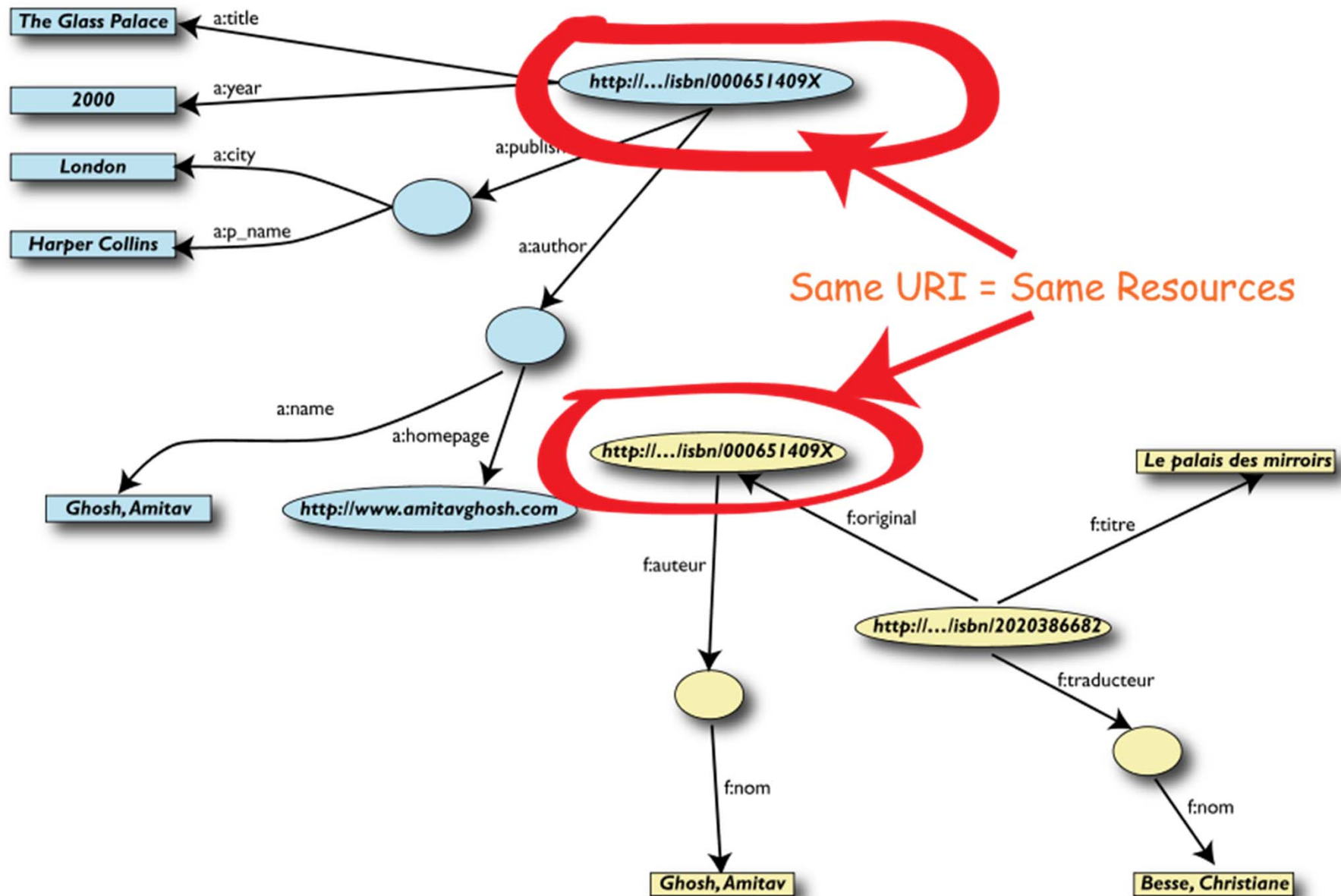
3rd: merge identical resources
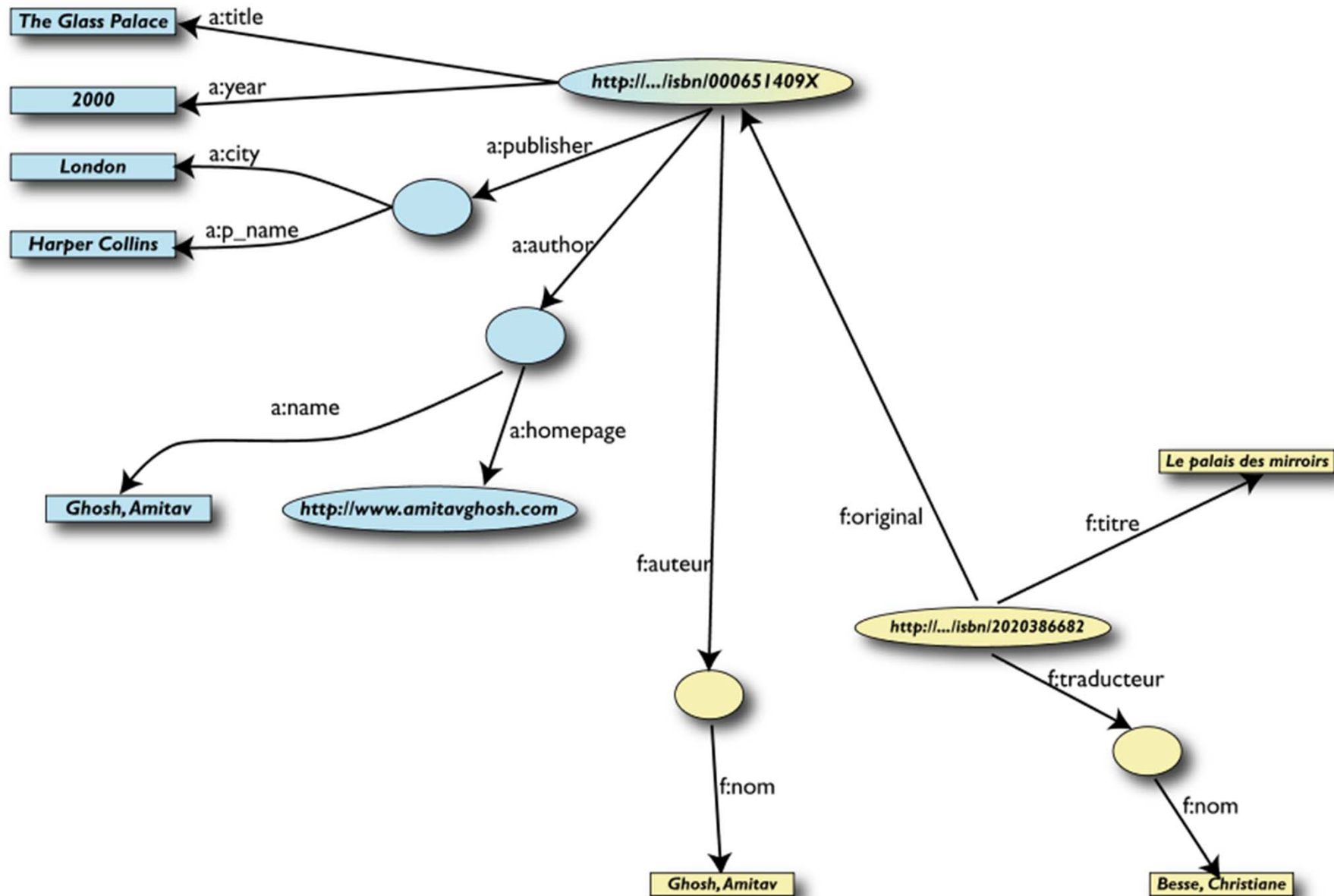
Start making queries…

- 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"!
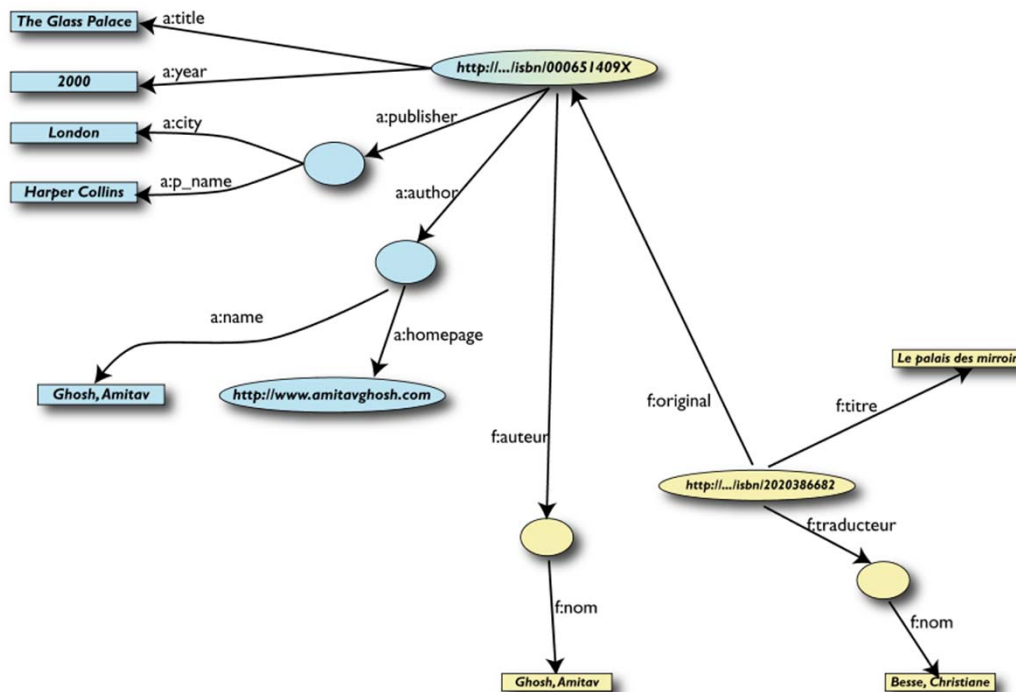
However, more can be achieved…

- 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
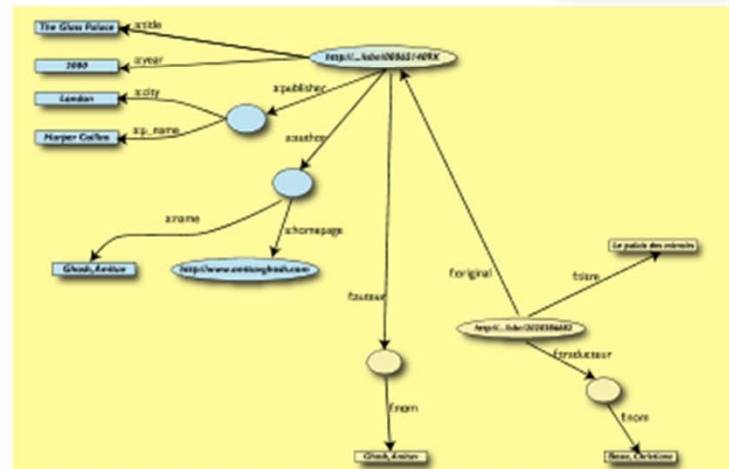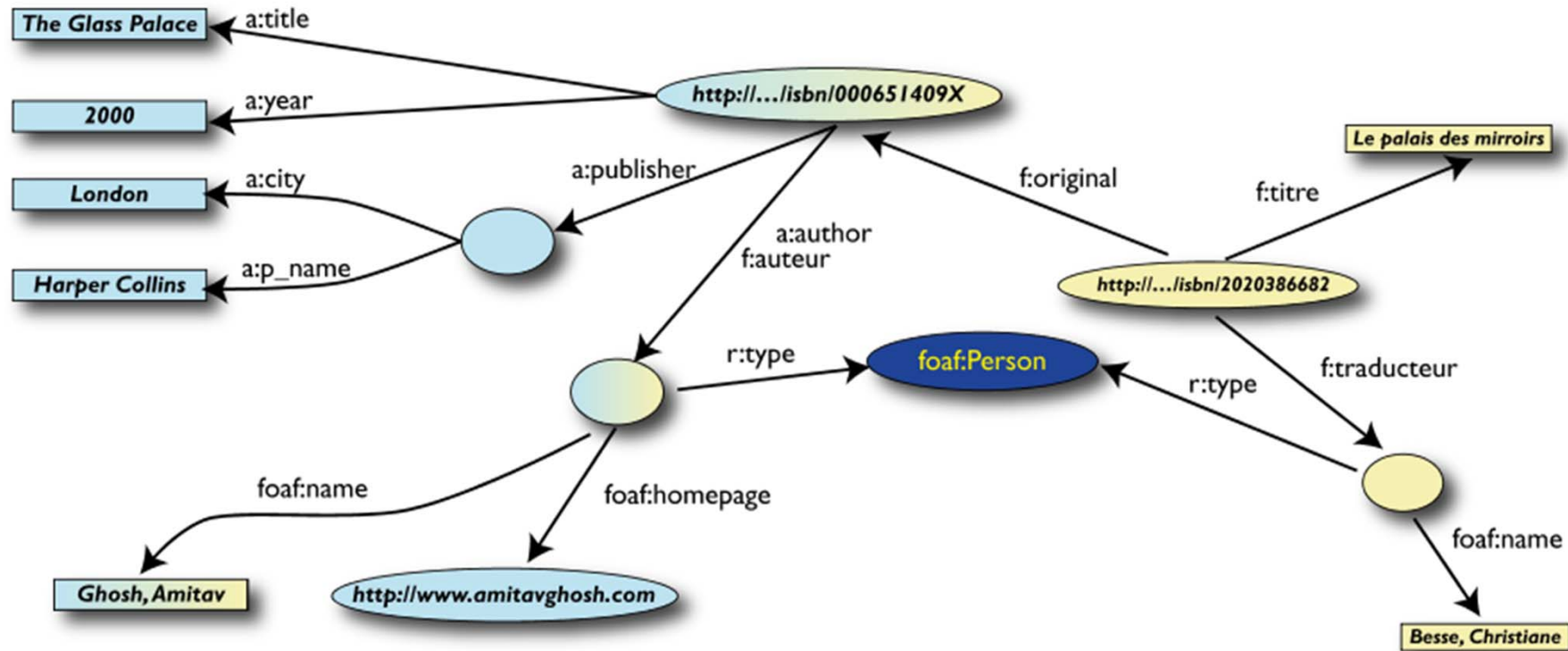
Start making richer queries!

- User of dataset "F" can now query:
  - □ "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"
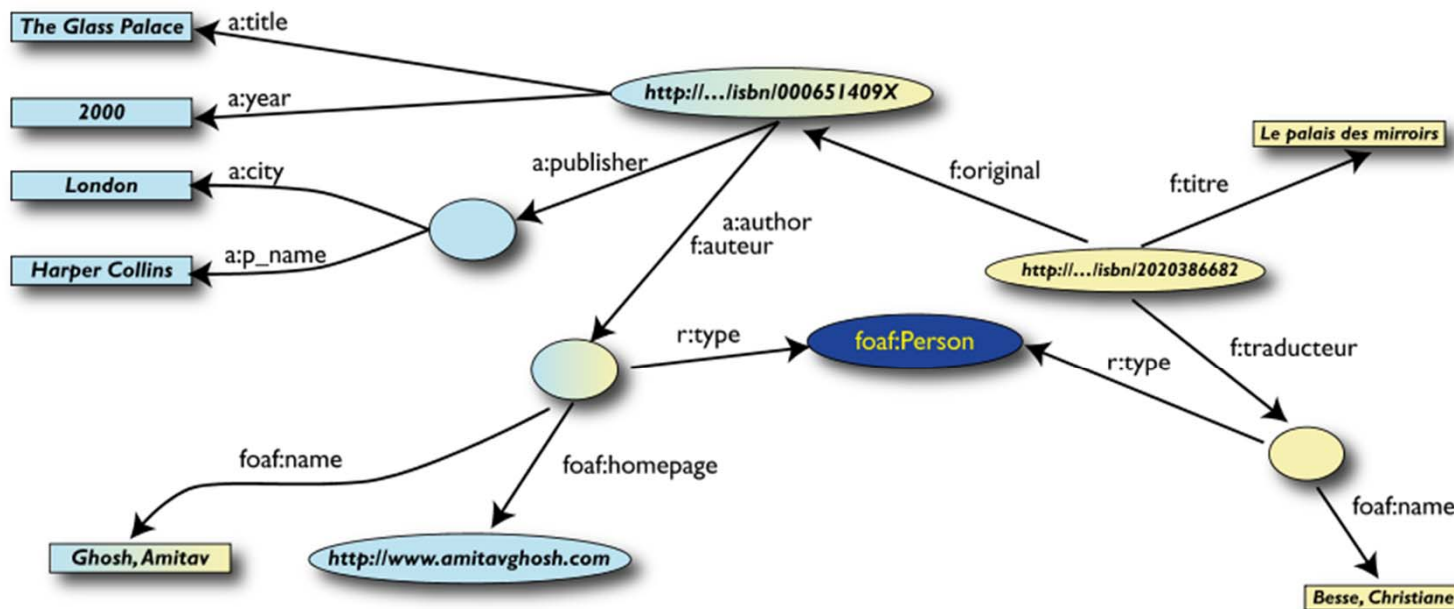
Combine with different datasets

- 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
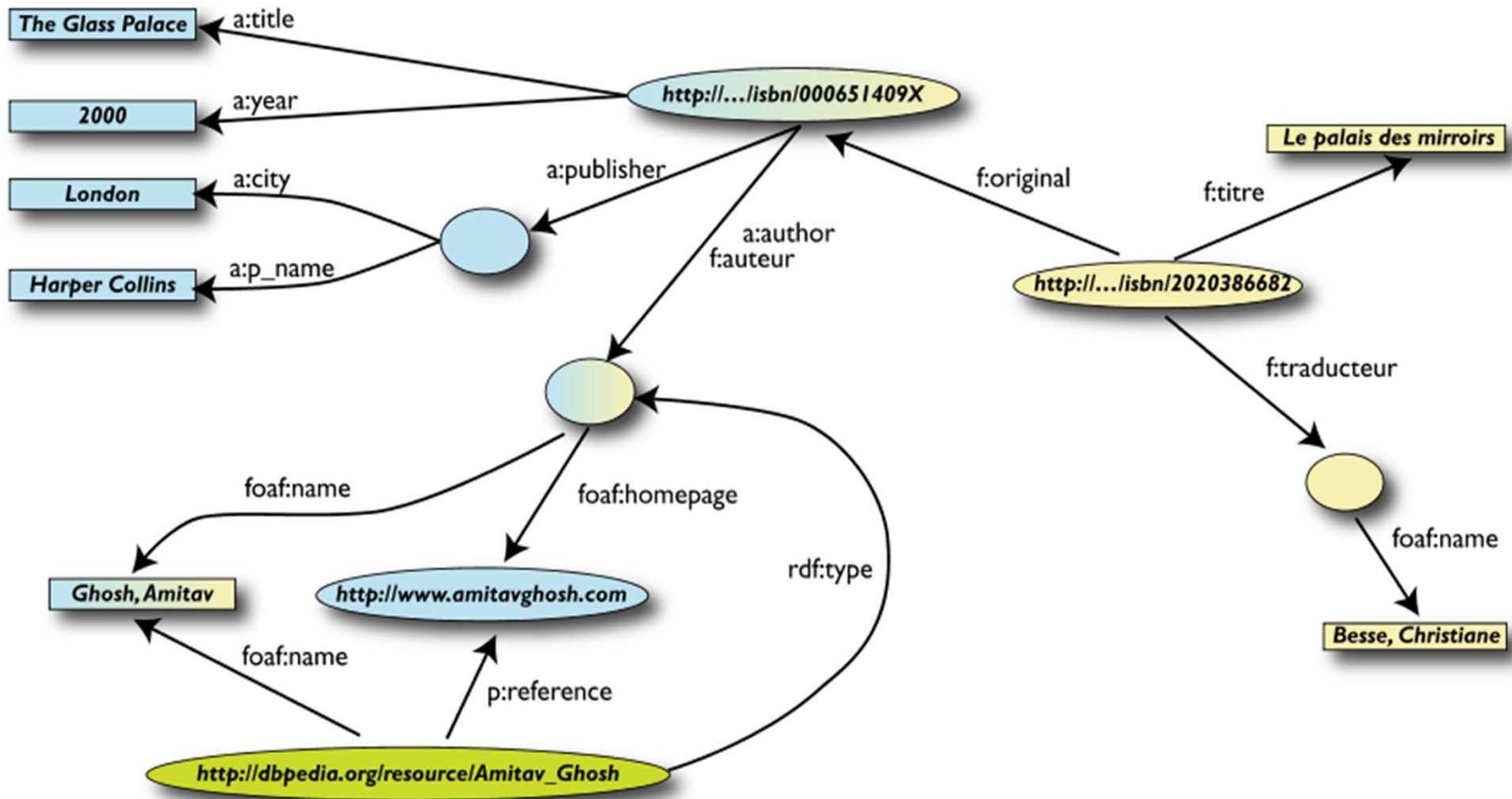
Merge with Wikipedia data
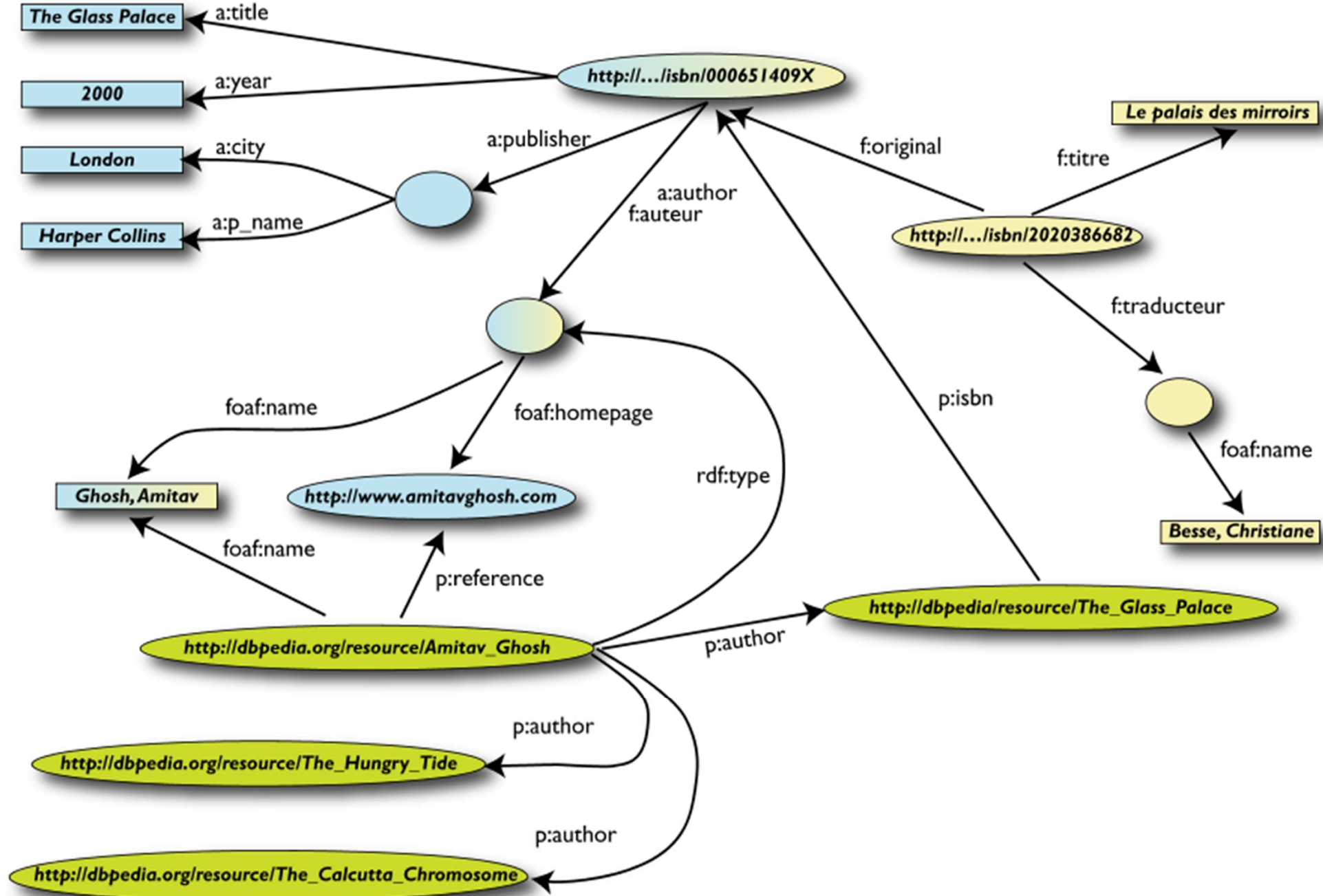
Merge with Wikipedia data
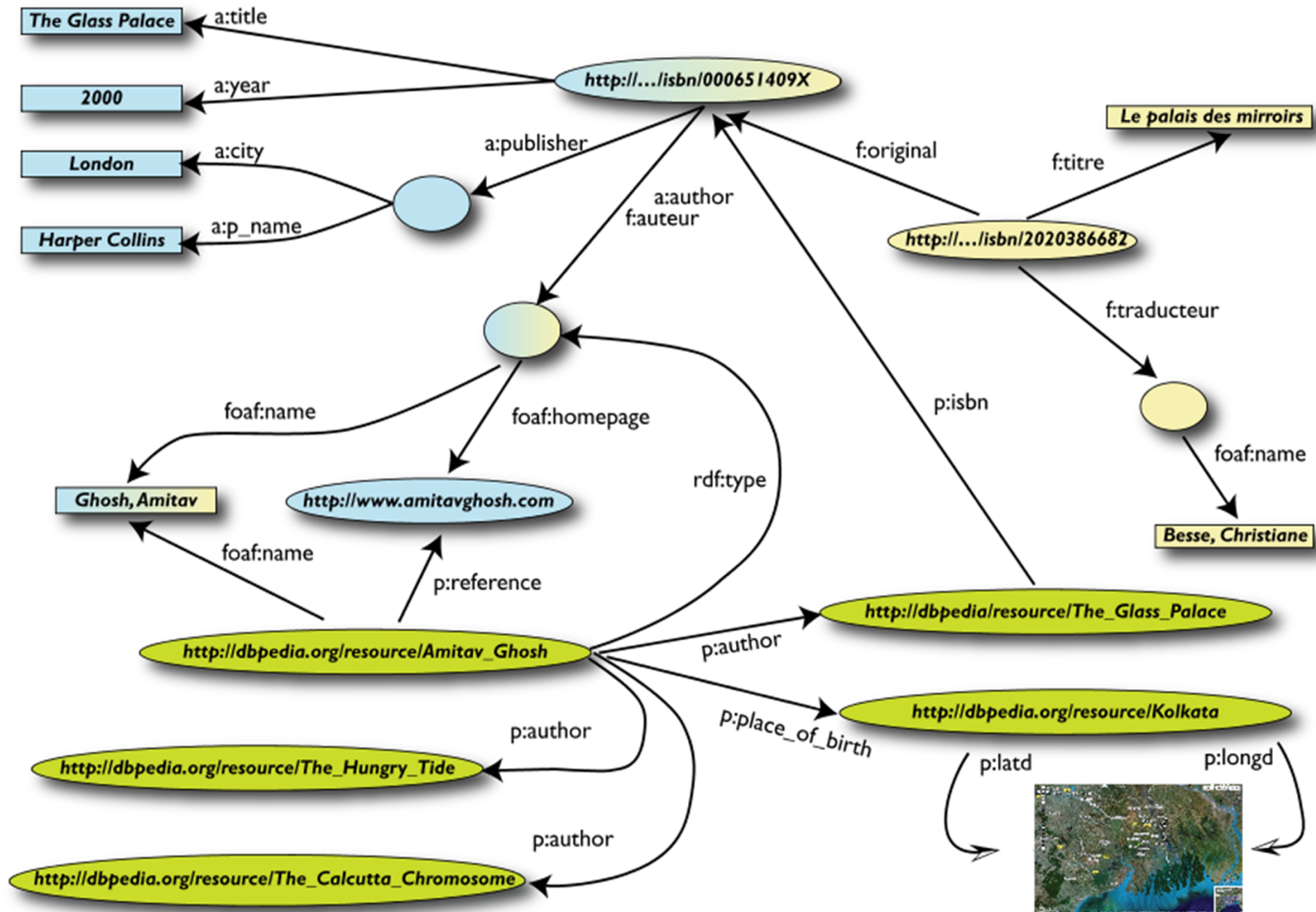
Is that surprising?

- 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, 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)

SPARQL

RDF
RDFS
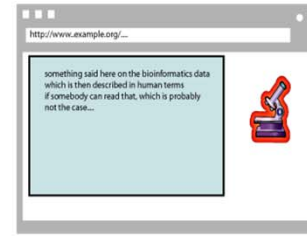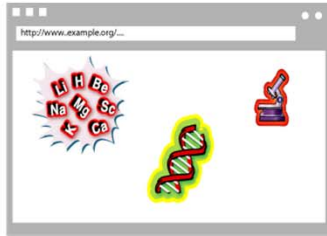
Applications

Query, Manipulate, etc.

Data represented in abstract format

Map, Expose, etc.

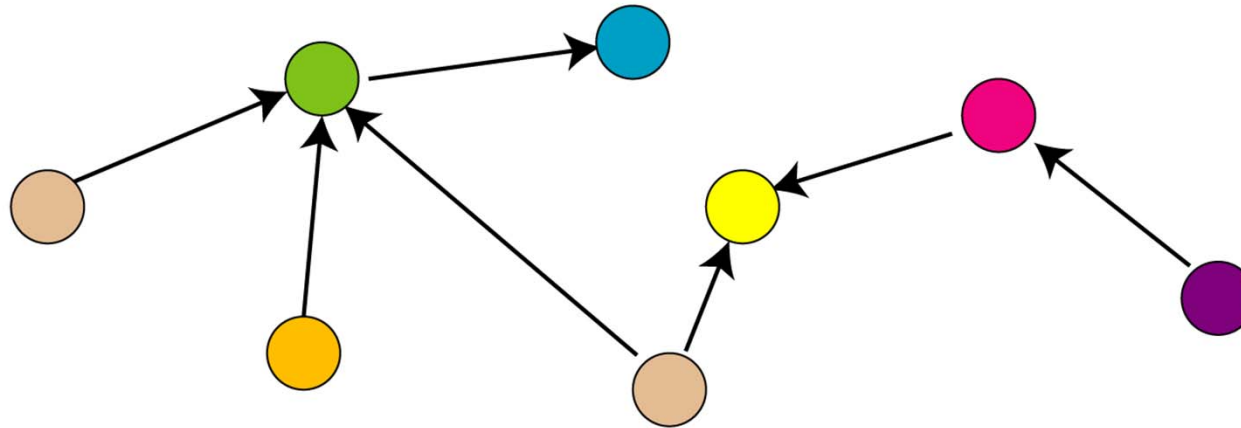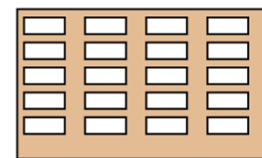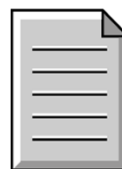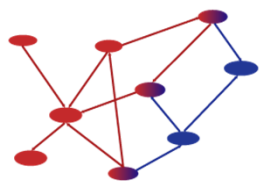Data in various formats

The Basis: RDF

- Let us begin to formalize what we did!
  - ☐ we "connected" the data…
  - ☐ but a simple connection is not enough… data should be named somehow
  - ☐ hence the RDF Triples: a labelled connection between two resources

Essential Online Resources

- http://www.w3.org/TR/html4/
- http://www.w3.org/Addressing/
- http://www.w3.org/Style/CSS/
- http://validator.w3.org/
- http://www.w3.org/