THE ROLE OF SEMANTICS IN DATA MANAGEMENT

AN INTRODUCTION TO SEMANTIC WEB, ONTOLOGIES & LINKED DATA

Outline

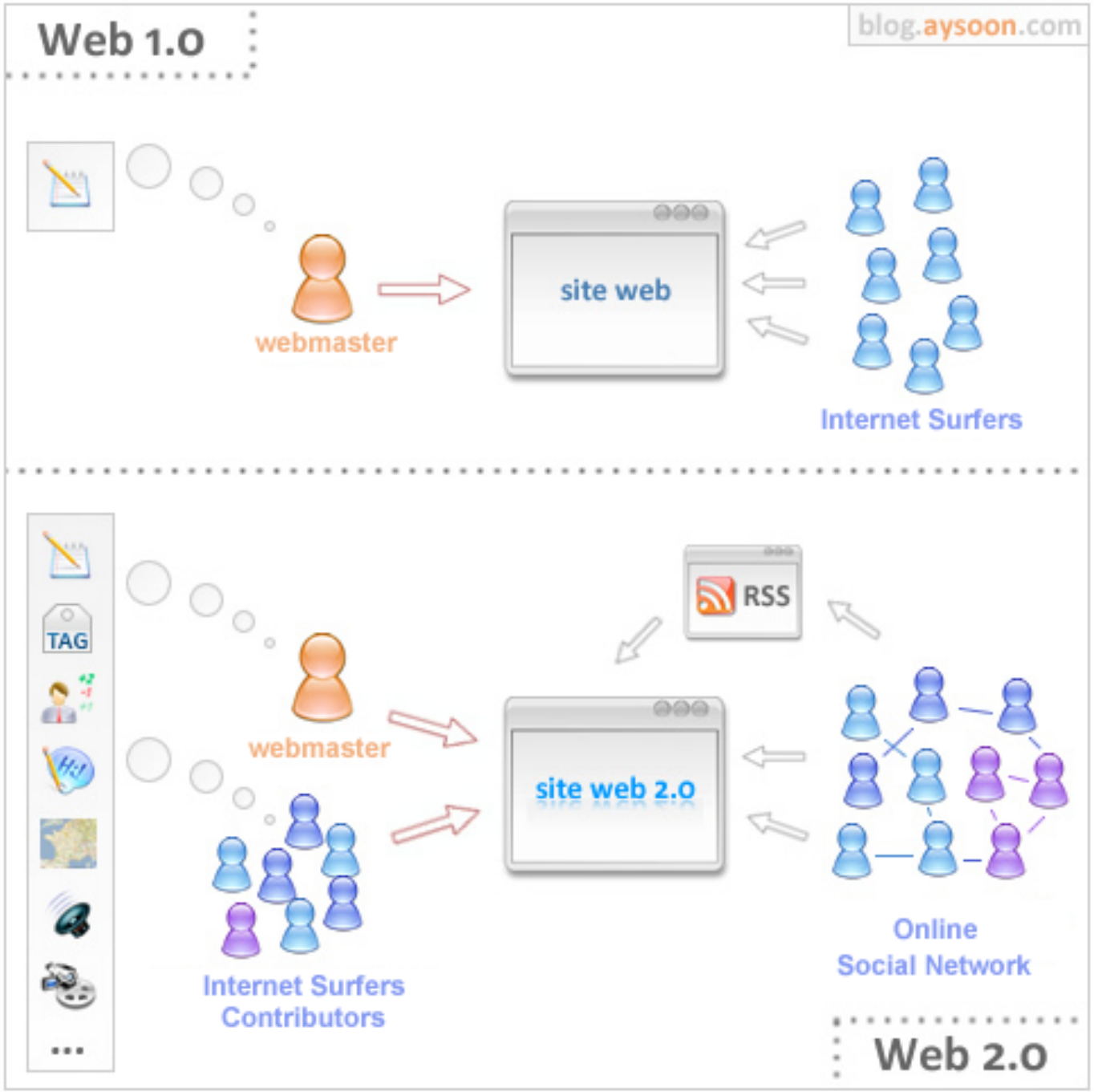
- From the Web to the Semantic Web (a.k.a. the Web of Data)
  - An a-priori semantics for data
- Ontologies
- Semantic Web approach
- Semantic Web technology
- Linked (Open) Data

Let’s start from the Web (1.0)

- The Web is a system of interlinked documents, accessed via Internet
- You can view Web pages and use hyperlinks to navigate between them
- People can easily access any of these documents
- This is the largest source of information ever
- *This is Web 1.0 or the Read-only Web*

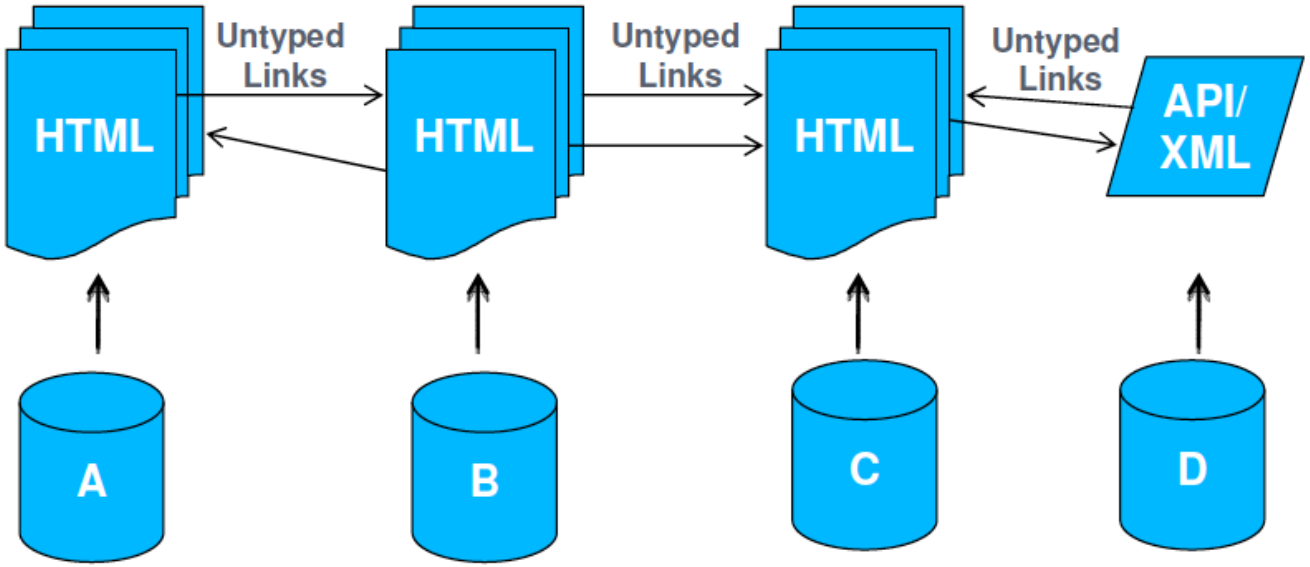
Then, we moved to Web 2.0

- Authors started to use Ajax
- People started to share multimedia content (photos and videos)
- People started to interact on social networks
- People started to publish content in blogs
- People started to contribute to wikis
- People started to use tags and RSS
- *This is Web 2.0 or the Read-Write Web*



Web 1.0 and Web 2.0

Web of Documents



Primary objects: **documents**

Links between **documents** (or parts of them)

Degree of structure in data: fairly **low**

Implicit semantics of contents

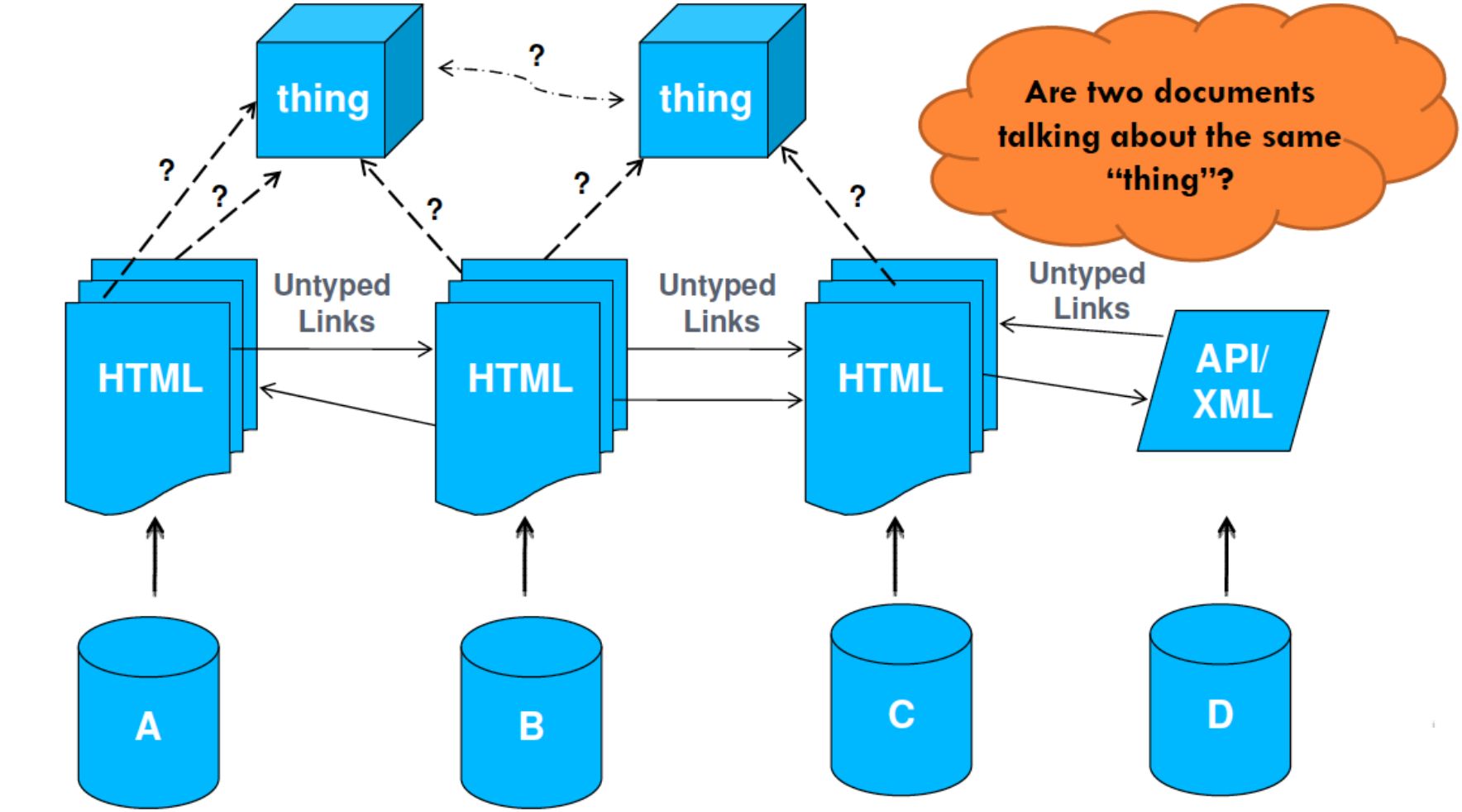
Designed for: **human consumption**

Based on presentations by Chris Bizer, Richard Cyganiak, Tom Heath, available at http://linkeddata.org/guides-and-tutorials

The problem

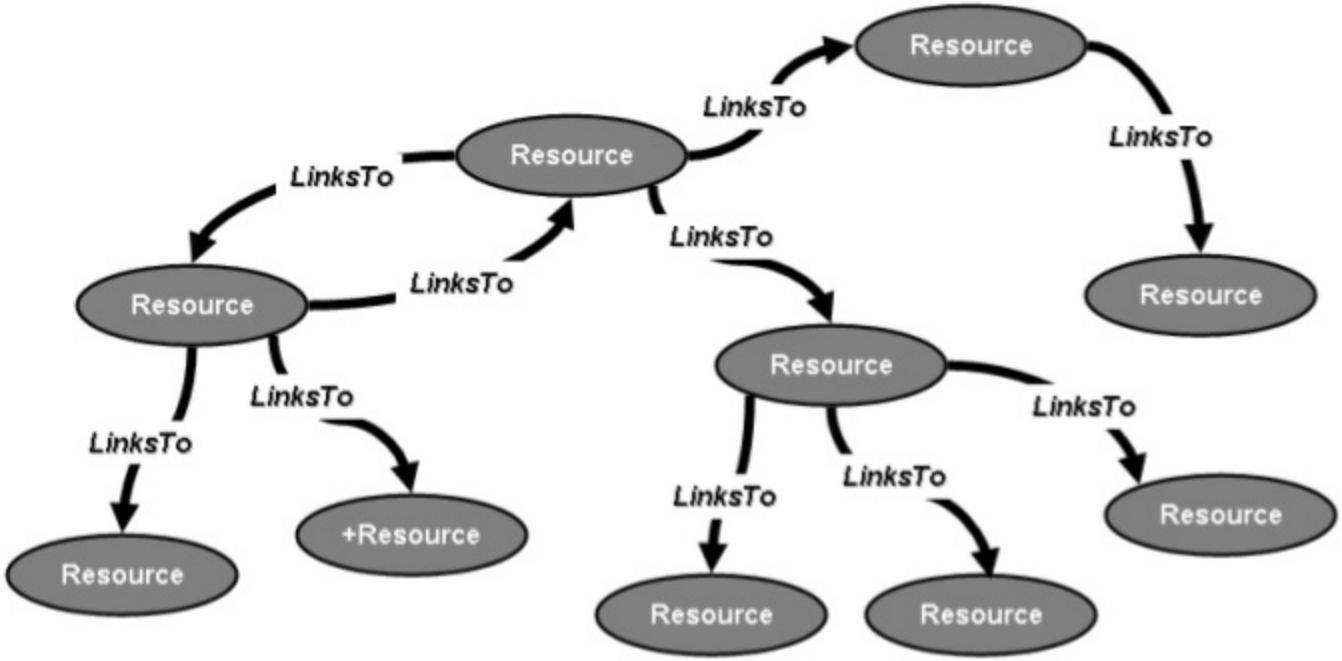
- Web pages are written in HTML
- HTML describes the *structure* of information
- HTML describes the *syntax* not the *semantics*
- if computers can understand the meaning behind information
  - they can learn what we are interested in
  - they can help us in better finding what we want
- we need to describe the *semantics* and find precise answers to complex questions
- this is what the *Semantic Web* is about
- this is Web 3.0:

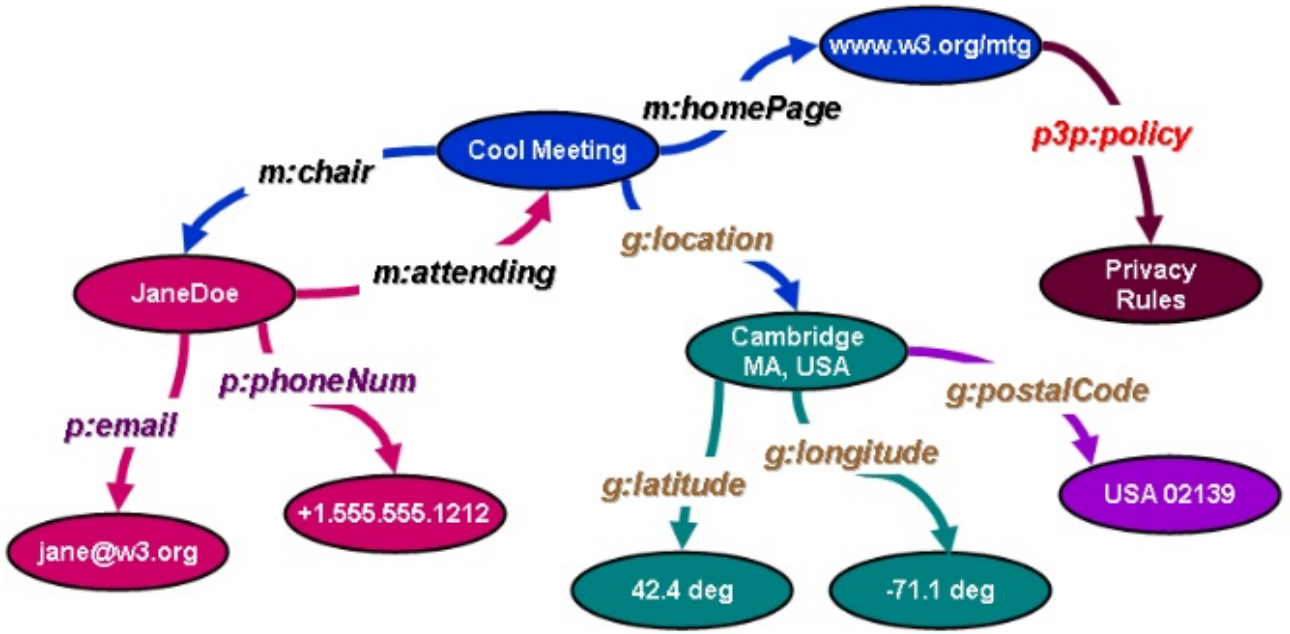
The problem



Based on presentations by Chris Bizer, Richard Cyganiak, Tom Heath, available at http://linkeddata.org/guides-and-tutorials

the Web is about documents





the Semantic Web is about things:

- it can recognize entities
- it can understand the relationship between things

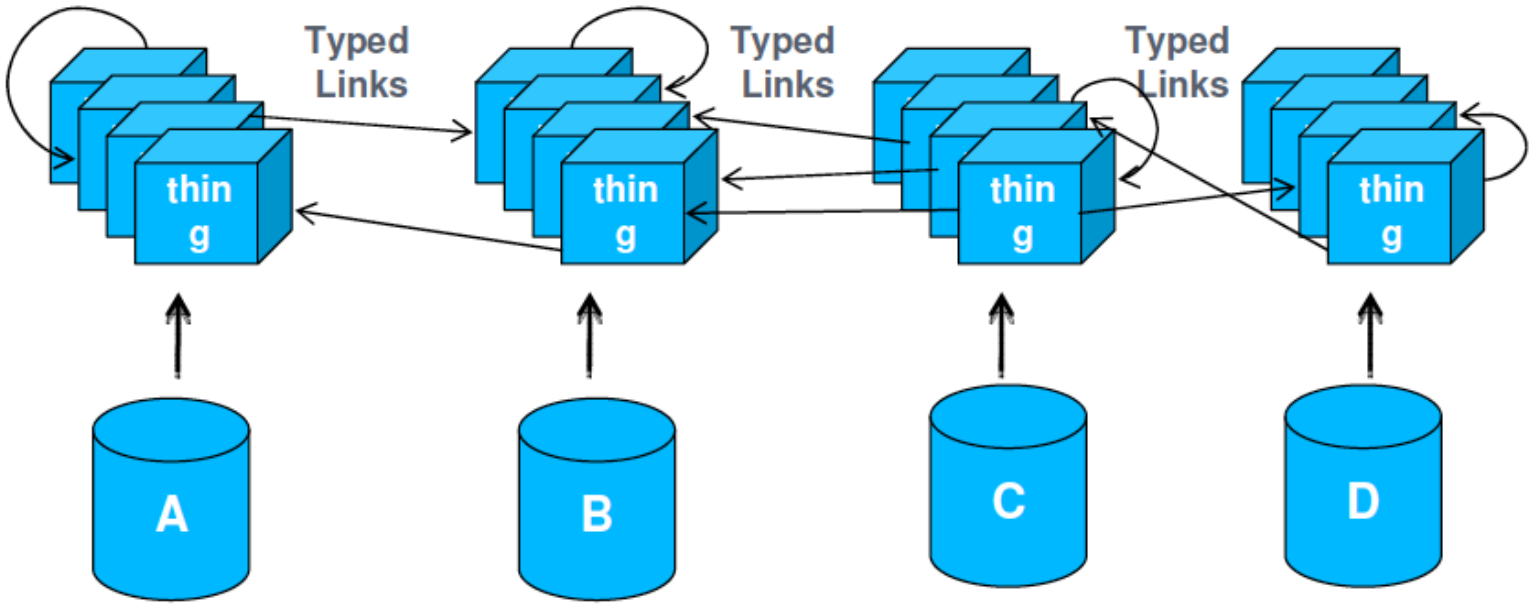
How

- Ex-post approach
  - information analysis
  - web scraping
  - natural language processing ➔ see NLP course
- Expensive!
  - human intervention
  - hard to maintain

How

- A priori approach
  - embedding semantic annotations into data
  - we need a *data model* for describing «things» (resources) and their interrelations
  - we need some languages to
    - *specify things and their interrelations* and
    - *perform reasoning* about them

Semantic Web a.k.a. Web of Data



Primary objects: **"things"** (or description of things)

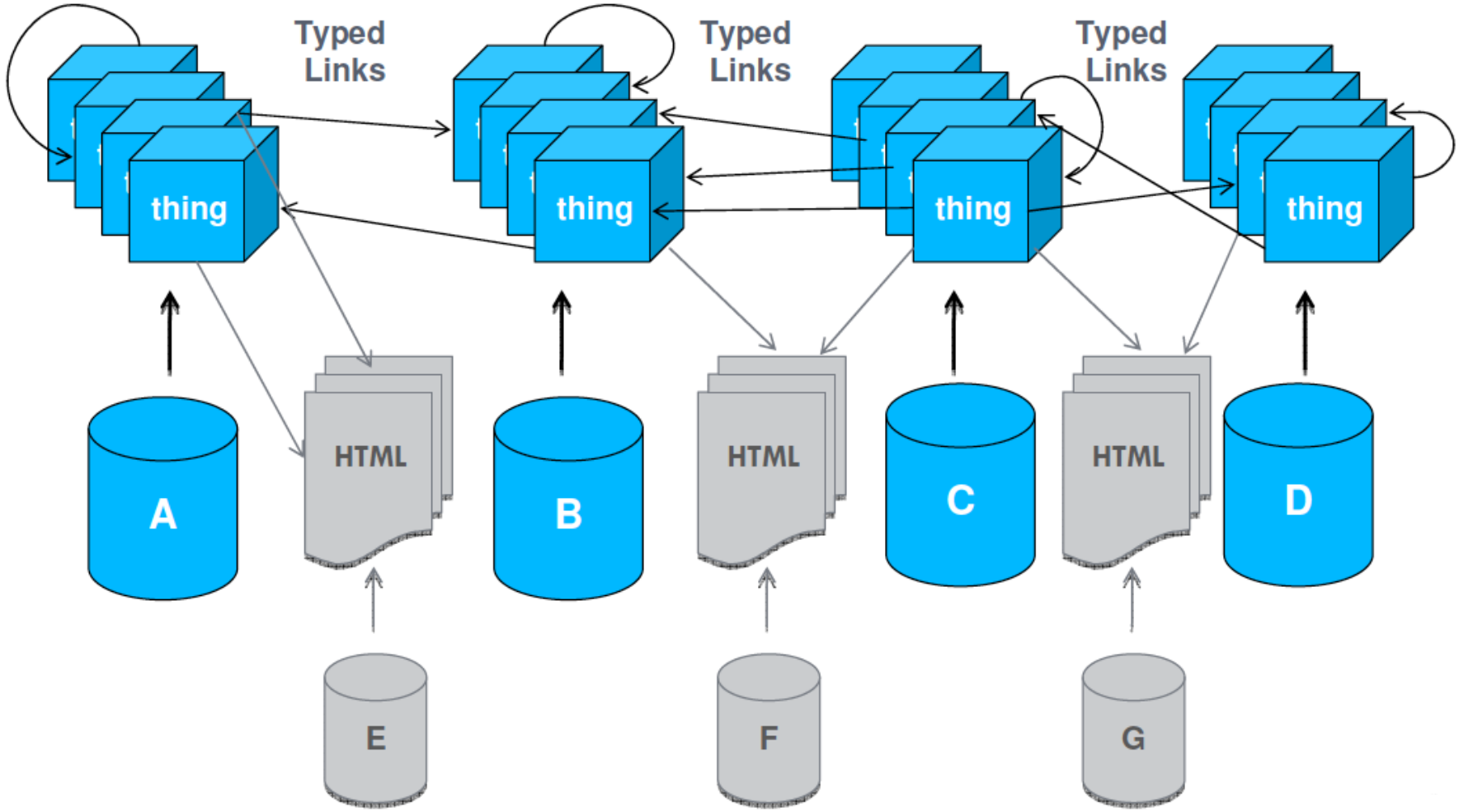
Links between **"things"**

Degree of Structure: **High**

Explicit semantics of contents **and links**

Designed for: Both **machines and humans**

Semantic Web a.k.a. Web of Data



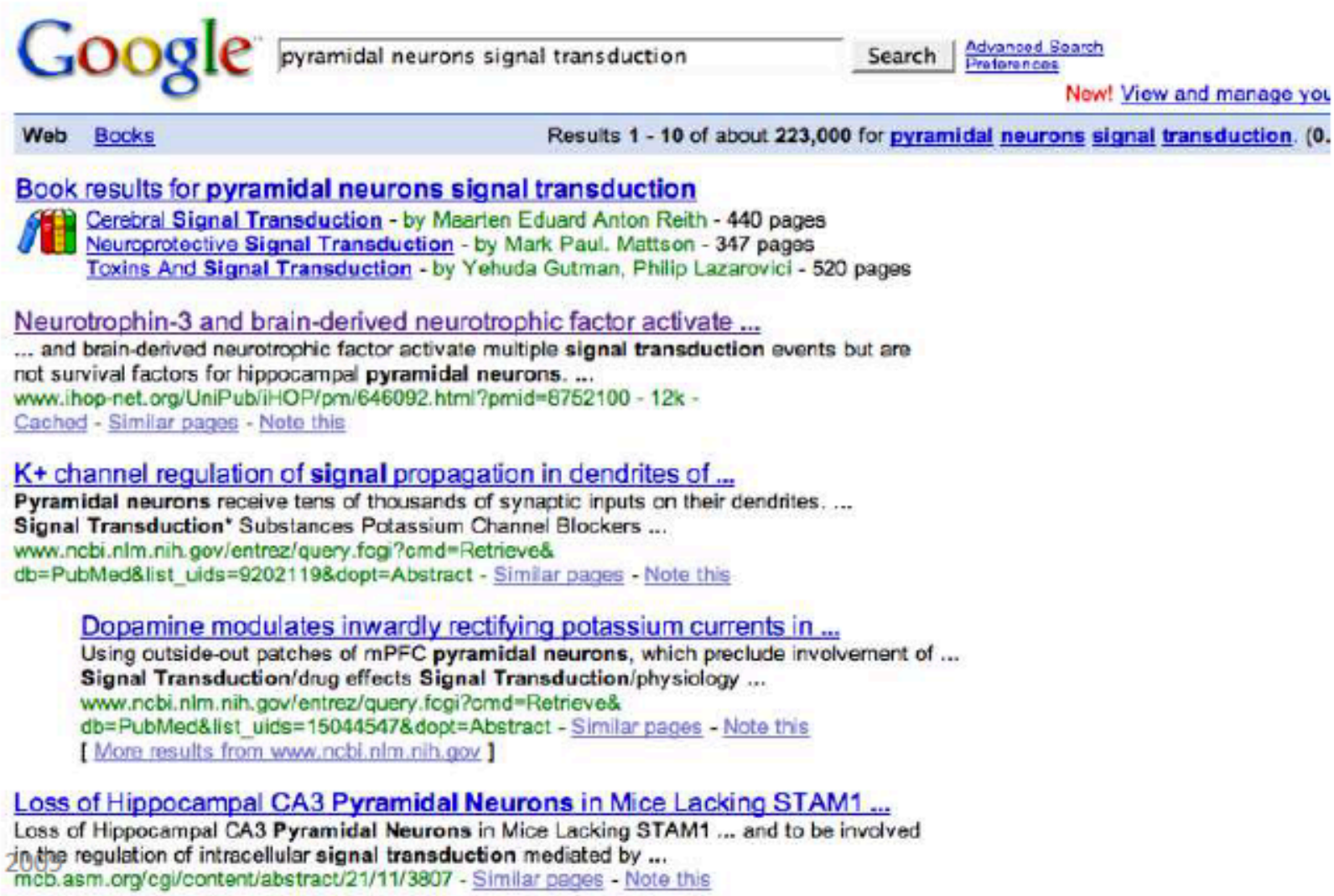
Based on presentations by Chris Bizer, Richard Cyganiak, Tom Heath, available at http://linkeddata.org/guides-and-tutorials

An example

Find me genes involved in signal transduction that are related to pyramidal neurons.

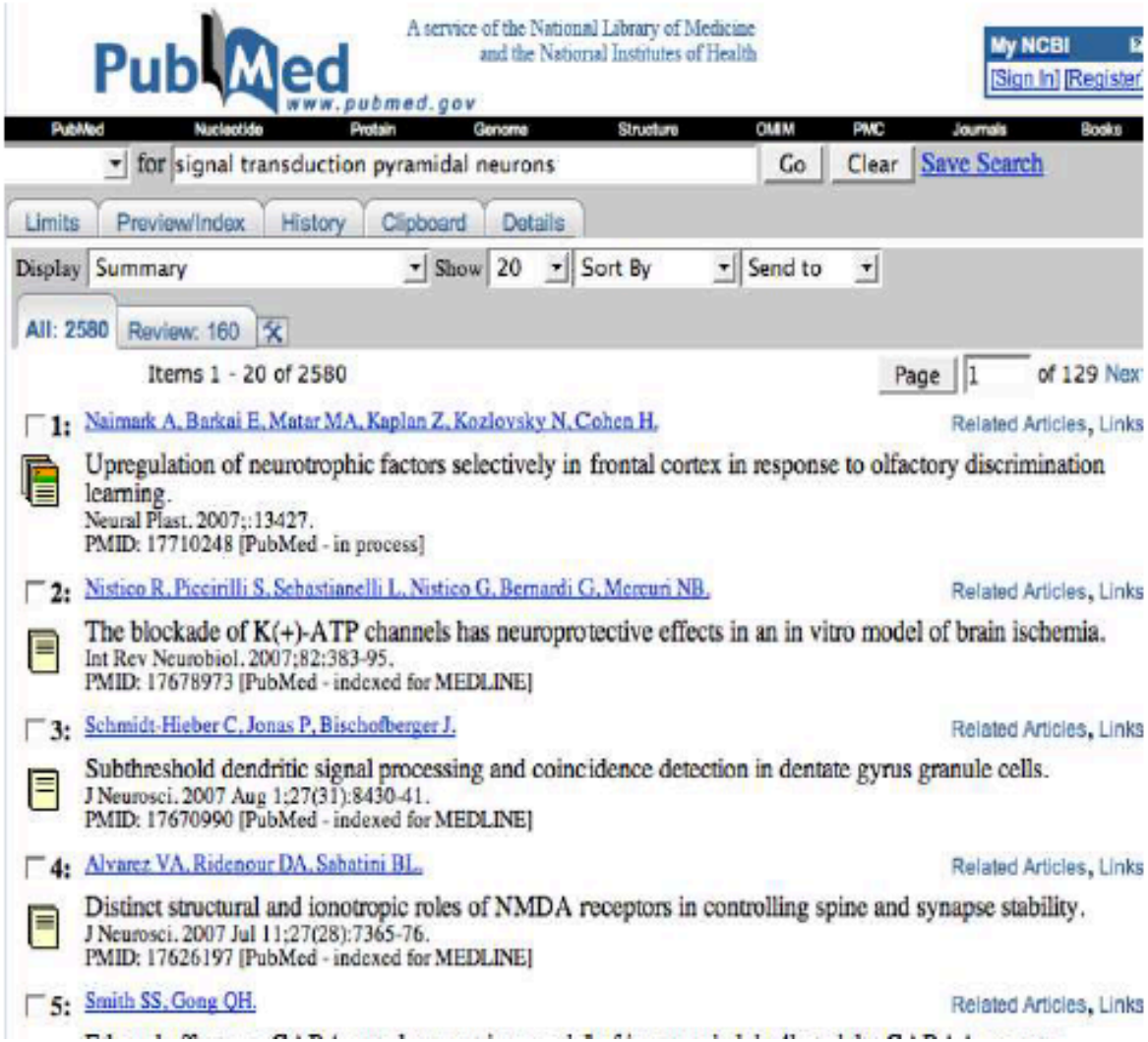
A general search

223,000 hits, 0 results



Domain-limited search

2,580 potential results



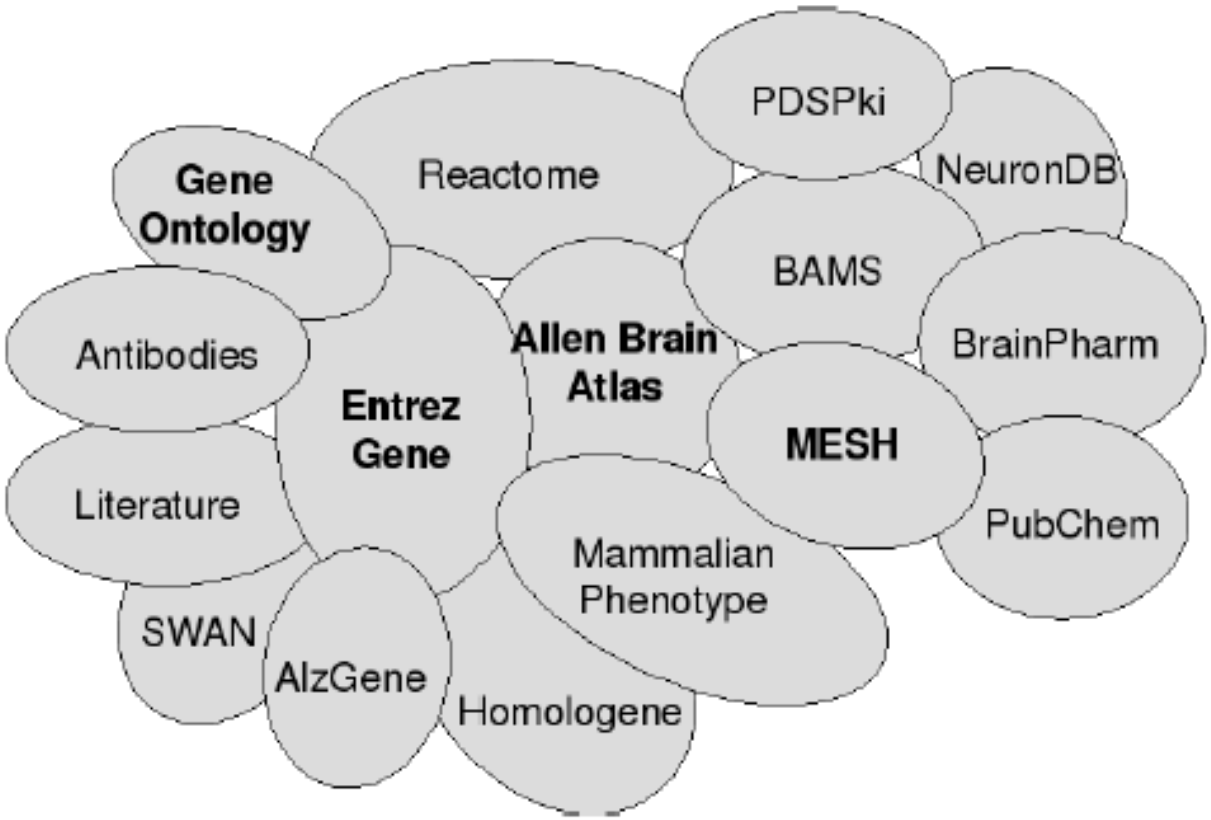
Specific databases

Too many silos!

A Semantic Web approach

Integrate disparate databases…

- MeSH
- PubMed
- Entrez Gene
- Gene Ontology
- …



A Semantic Web Approach

...so that *one* query...

```
prefix go: <http://purl.org/obo/owl/GO#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix mesh: <http://purl.org/commons/record/mesh/>
prefix sc: <http://purl.org/science/owl/sciencecommons/>
prefix ro: <http://www.obofoundry.org/ro/ro.owl#>
prefix senselab: <http://purl.org/ycmi/senselab/neuron_ontology.owl#>

SELECT ?genename ?processname ?receptor_protein_name
 WHERE {
  # PubMeSH includes ?gene_records mentioned in ?articles which are identified by pmid in ?pubmed_records.
  GRAPH <http://purl.org/commons/hcls/pubmesh> {
    ?pubmed_record sc:has-as-minor-mesh mesh:D017966 .
    ?article sc:identified_by_pmid ?pubmed_record .
    ?gene_record sc:describes_gene_or_gene_product_mentioned_by ?article
  }
   # The Gene Ontology asserts that foreach ?protein, ?protein ro:has_function [ ro:realized_as ?process ].
  GRAPH <http://purl.org/commons/hcls/goa> {
   ?protein rdfs:subClassOf ?restriction1 .
   ?restriction1 owl:onProperty ro:has_function .
   ?restriction1 owl:someValuesFrom ?restriction2 .
   ?restriction2 owl:onProperty ro:realized_as .
   ?restriction2 owl:someValuesFrom ?process .
  # Also, foreach ?protein, ?protein has a parent class which is linked by some predicate to ?gene_record.
   ?protein rdfs:subClassOf ?protein_superclass .
   ?protein_superclass owl:equivalentClass ?restriction3 .
   ?restriction3 owl:onProperty sc:is_protein_gene_product_of_dna_described_by .
   ?restriction3 owl:hasValue ?gene_record .
  # Each ?process (that we are interested in) is a subclass of the signal transduction process.
  GRAPH <http://purl.org/commons/hcls/20070416/classrelations> {
    { ?process obo:part_of go:GO_0007166 }
    UNION
    { ?process rdfs:subClassOf go:GO_0007166 }
  } }
  GRAPH <http://purl.org/commons/hcls/gene> {
   ?gene_record rdfs:label ?genename
  }
  GRAPH <http://purl.org/commons/hcls/20070416> {
   ?process rdfs:label ?processname
```

Mesh: Pyramidal Neurons

↓

Pubmed: Journal Articles

↓

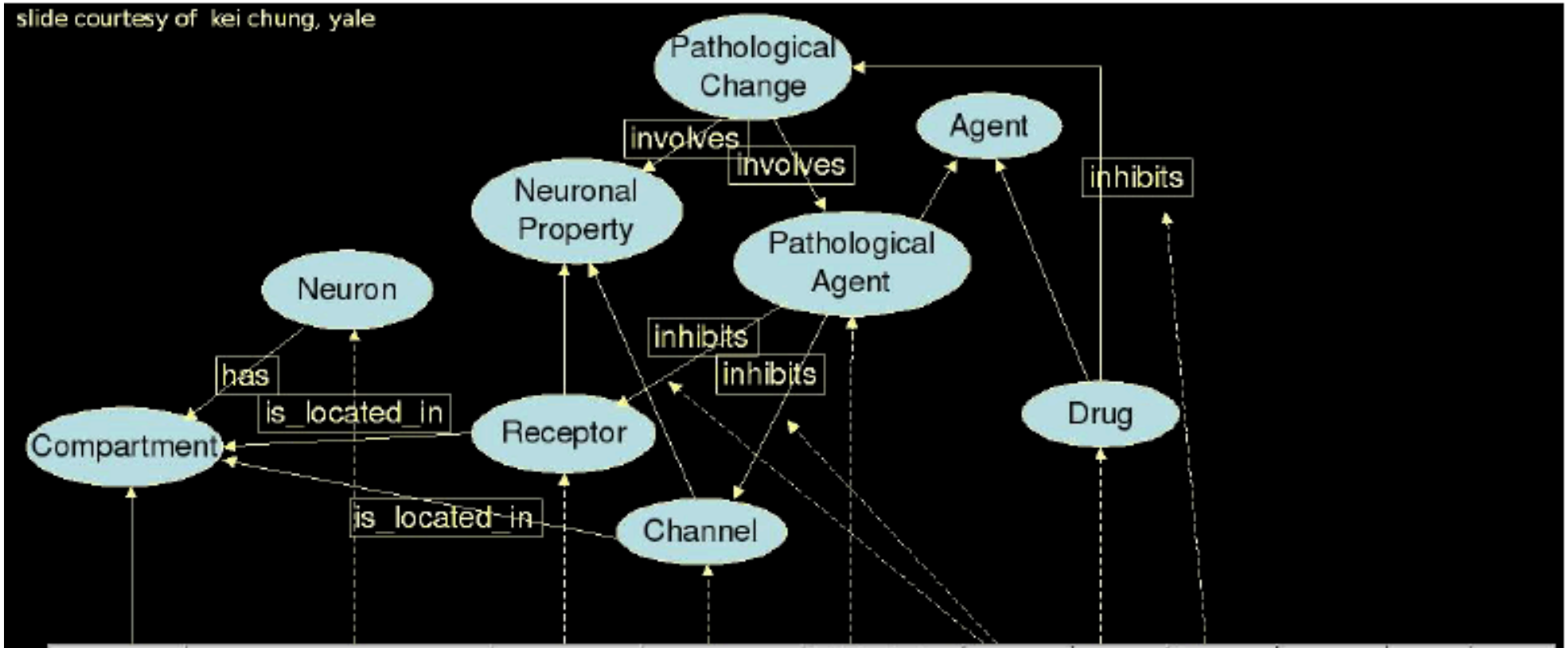
Entrez Gene: Genes

↓

GO: Signal Transduction

A Semantic Web Approach

…(trivially) spans several databases…



Compartment	Cell: NeuronDB	Receptor	Channel	Pathological Agent (PA)	PA Action	Drug	Drug Action	Stage	Note	Detail
Soma	CA1 pyramidal neuron		I A	beta Amyloid	Inhibits			Early	View	66240
	Olfactory bulb mitral cell	GabaA						Early	View	66250
Dendrite	CA1 pyramidal neuron		I A	beta Amyloid	Inhibits			Early	View	66240
	Olfactory bulb mitral cell	GabaA						Early	View	66250
Unspecified	Oocyte		I L high threshold	beta Amyloid	Inhibits			Early	View	66252
								Early	View	66253
	CA1 pyramidal neuron			beta Amyloid	Inhibits			Early	View	66258
	CA1 pyramidal neuron	NMDA	I Calcium	beta Amyloid	Inhibits		Inhibits		View	66250

A Semantic Web Approach

…to deliver targeted results…

DRD1, 1812	adenylate cyclase activation
ADRB2, 154	adenylate cyclase activation
ADRB2, 154	arrestin mediated desensitization of G-protein coupled receptor protein signaling pathway
DRD1IP, 50632	dopamine receptor signaling pathway
DRD1, 1812	dopamine receptor, adenylate cyclase activating pathway
DRD2, 1813	dopamine receptor, adenylate cyclase inhibiting pathway
GRM7, 2917	G-protein coupled receptor protein signaling pathway
GNG3, 2785	G-protein coupled receptor protein signaling pathway
GNG12, 55970	G-protein coupled receptor protein signaling pathway
DRD2, 1813	G-protein coupled receptor protein signaling pathway
ADRB2, 154	G-protein coupled receptor protein signaling pathway
CALM3, 808	G-protein coupled receptor protein signaling pathway
HTR2A, 3356	G-protein coupled receptor protein signaling pathway
DRD1, 1812	G-protein signaling, coupled to cyclic nucleotide second messenger
SSTR5, 6755	G-protein signaling, coupled to cyclic nucleotide second messenger
MTNR1A, 4543	G-protein signaling, coupled to cyclic nucleotide second messenger
CNR2, 1269	G-protein signaling, coupled to cyclic nucleotide second messenger
HTR6, 3362	G-protein signaling, coupled to cyclic nucleotide second messenger
GRIK2, 2898	glutamate signaling pathway
GRIN1, 2902	glutamate signaling pathway
GRIN2A, 2903	glutamate signaling pathway
GRIN2B, 2904	glutamate signaling pathway
ADAM10, 102	integrin-mediated signaling pathway
GRM7, 2917	negative regulation of adenylate cyclase activity
LRP1, 4035	negative regulation of Wnt receptor signaling pathway
ADAM10, 102	Notch receptor processing
ASCL1, 429	Notch signaling pathway
HTR2A, 3356	serotonin receptor signaling pathway
ADRB2, 154	transmembrane receptor protein tyrosine kinase activation (dimerization)
PTPRG, 5793	transmembrane receptor protein tyrosine kinase signaling pathway
EPHA4, 2043	transmembrane receptor protein tyrosine kinase signaling pathway
NRTN, 4902	transmembrane receptor protein tyrosine kinase signaling pathway

What's the trick?

- Agreement on common terms and relationships
- Incremental, flexible data structure
- Good-enough modeling
- Query interface tailored to the data model

- key concept: ontology for the Web

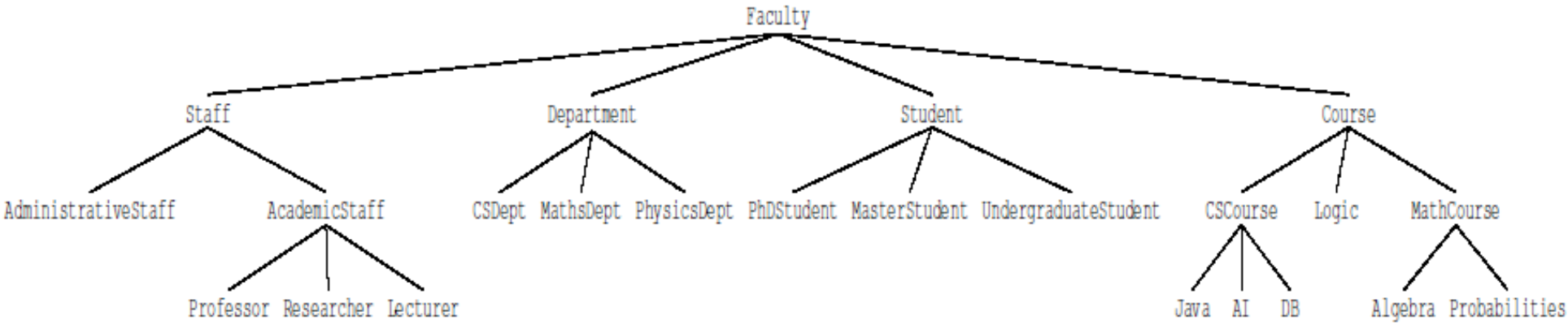
Ontologies

Ontologies

- Formal descriptions providing human users a shared understanding of a given domain
  - A controlled vocabulary
- Formally defined so that it can also be processed by machines
- Logical semantics that enables reasoning.
- Reasoning is the key for different important tasks of Web data management, in particular
  - to answer queries (over possibly distributed data)
  - to relate objects in different data sources enabling their integration
  - to detect inconsistencies or redundancies
  - to refine queries with too many answers, or to relax queries with no answer

Classes and Class Hierarchies

- Backbone of the ontology
- AcademicStaff is a Class
- (A class will be interpreted as a set of objects)
- AcademicStaff isa Staff
- (isa is interpreted as set inclusion)



Relations

- Declaration of relations with their signature
- (Relations will be interpreted as binary relations between objects)
- `TeachesIn(AcademicStaff, Course)`
  - if one states that "*X* `TeachesIn` *Y*", then *X* belongs to `AcademicStaff` and *Y* to `Course`,
- `TeachesTo(AcademicStaff, Student)`,
- `Leads(Staff, Department)`

Instances

- Classes have instances
- `Dupond` is an instance of the class `Professor`
- it corresponds to the fact: `Professor(Dupond)`

- Relations also have instances
- `(Dupond,CS101)` is an instance of the relation `TeachesIn`
- it corresponds to the fact: `TeachesIn(Dupond,CS101)`

- The instance statements can be seen as (and stored in) a database

Ontology = schema + instance

- Schema
  - The set of class and relation names
  - The signatures of relations and also constraints
  - The constraints that are used for two purposes
    - ★ checking data consistency (like dependencies in databases)
    - ★ inferring new facts
- Instance
  - The set of facts
  - The set of base facts together with the inferred facts should satisfy the constraints
- Ontology (i.e., Knowledge Base) = Schema + Instance

Reasoning

- The use of logic as ontology semantics enables reasoning
- For instance, from the following facts:

  - Leads(Dupond, CS Department)
  - Leads(Staff, Department)

  it follows that

  - Staff(Dupond)
  - Department(CS Department)

The Semantic Web

- Augments the World Wide Web through the usage of specific ontologies
- Represents the Web's information in a machine-readable fashion
- Enables…
  - … targeted search
  - … data browsing
  - … automated agents

Semantic Web technologies

- A family of technology standards that 'play nice together', including:
  - Flexible data model
  - Distributed query language
  - Expressive ontology language

The technologies enable us to build applications and solutions that were not possible, practical, or feasible traditionally.

The Semantic Web approach

How does a Semantic Web approach help us merge data sets, infer new relations, and integrate outside data sources?

Thanks to Ivan Herman for this example

The Semantic Web approach

- Map the various data onto an abstract data representation (instances and schema of some ontology)
- Make the data independent of its internal representation
- Merge the resulting representations
- Start making queries on the whole
  - Queries not possible on the individual data sets

Dataset “A”: a simplified book store

Books

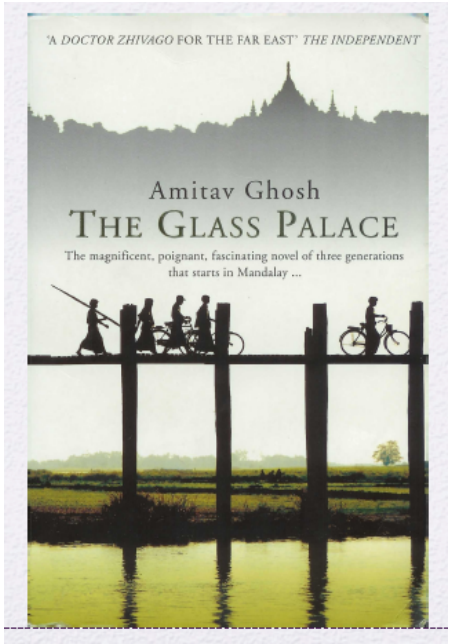
ID	Author	Title	Publisher	Year
ISBN0-00-651409-X	id_xyz	The Glass Palace	id_qpr	2000

Authors

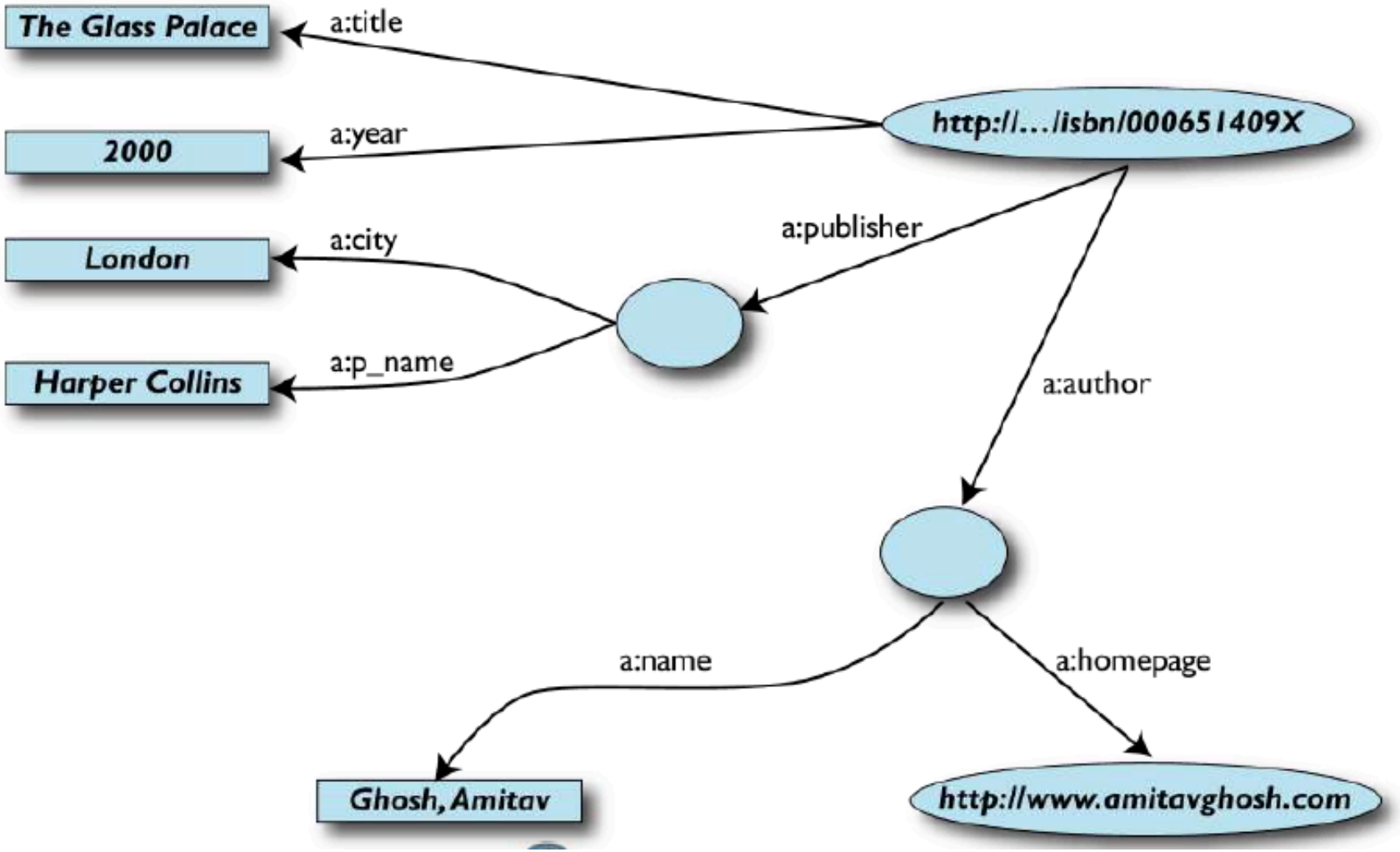
ID	Name	Home page
id_xyz	Ghosh, Amitav	http://www.amitavghosh.com

Publishers

ID	Publisher Name	City
id_qpr	Harper Collins	London



1st: export your data as a set of relations among entities (ontology instance level)

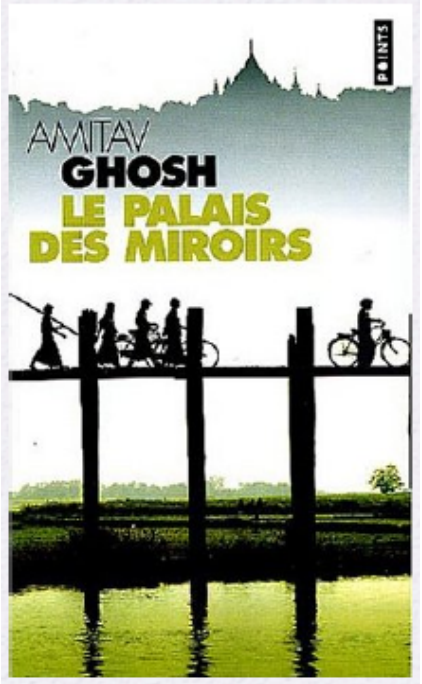


Dataset “F”: another book store

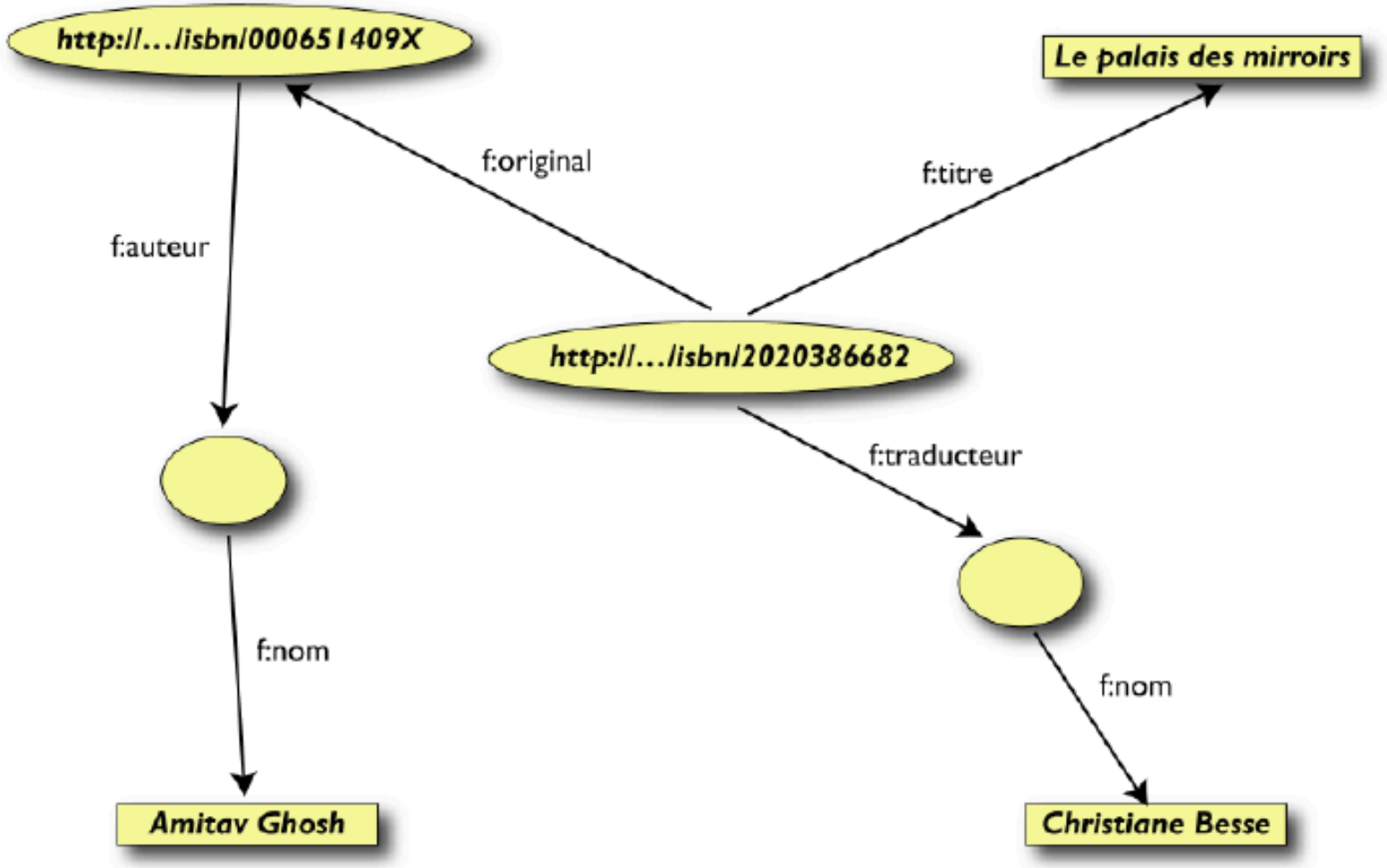
	A	B	D	E
1	**ID**	**Titre**	**Traducteur**	**Original**
2	ISBN0 2020386682	Le Palais des miroirs	A13	ISBN-0-00-651409-X
3				

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

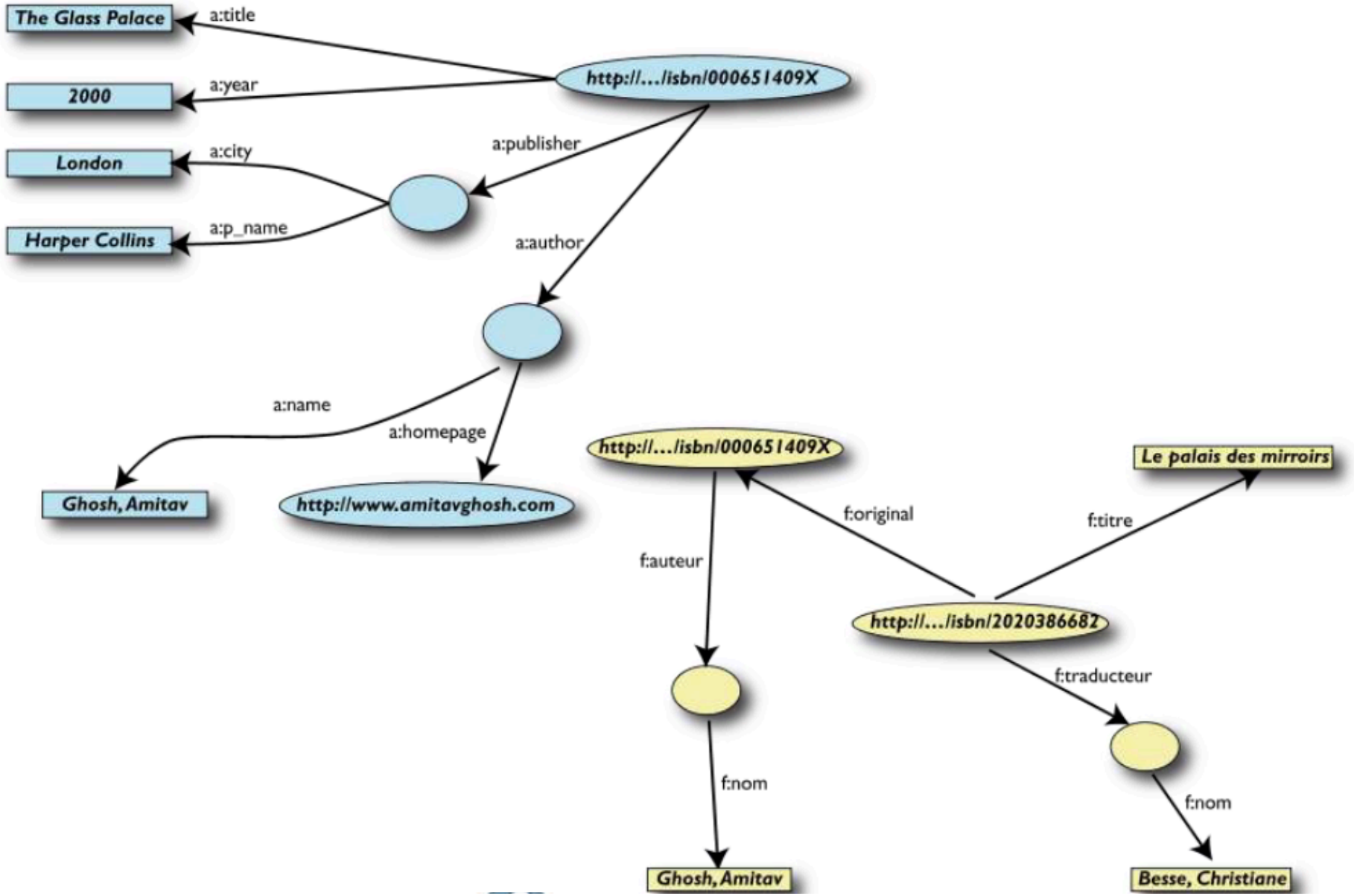
	Nom
11	**Nom**
12	Ghosh, Amitav
13	Besse, Christianne



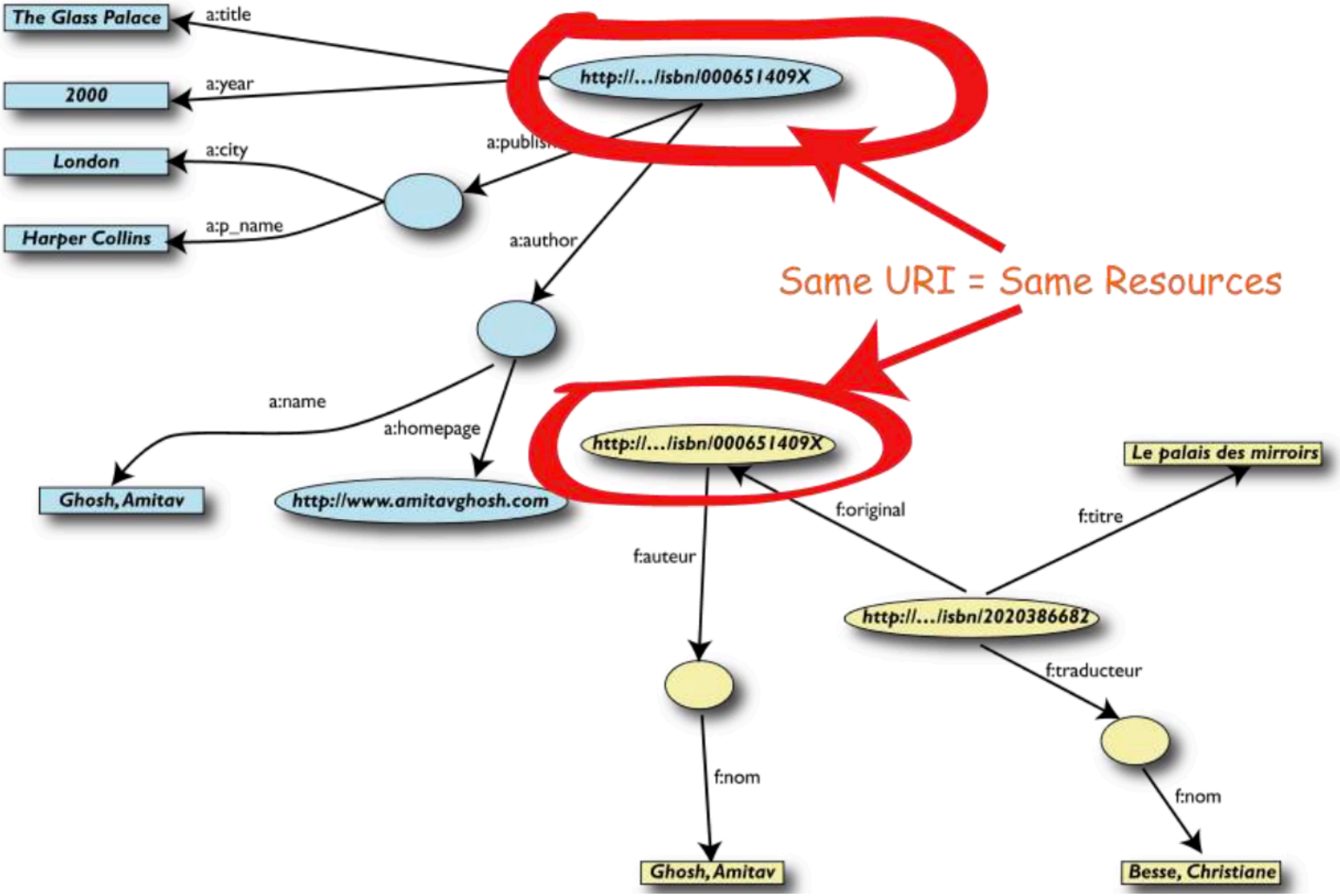
2nd: export your second set of data (ontology instance level)



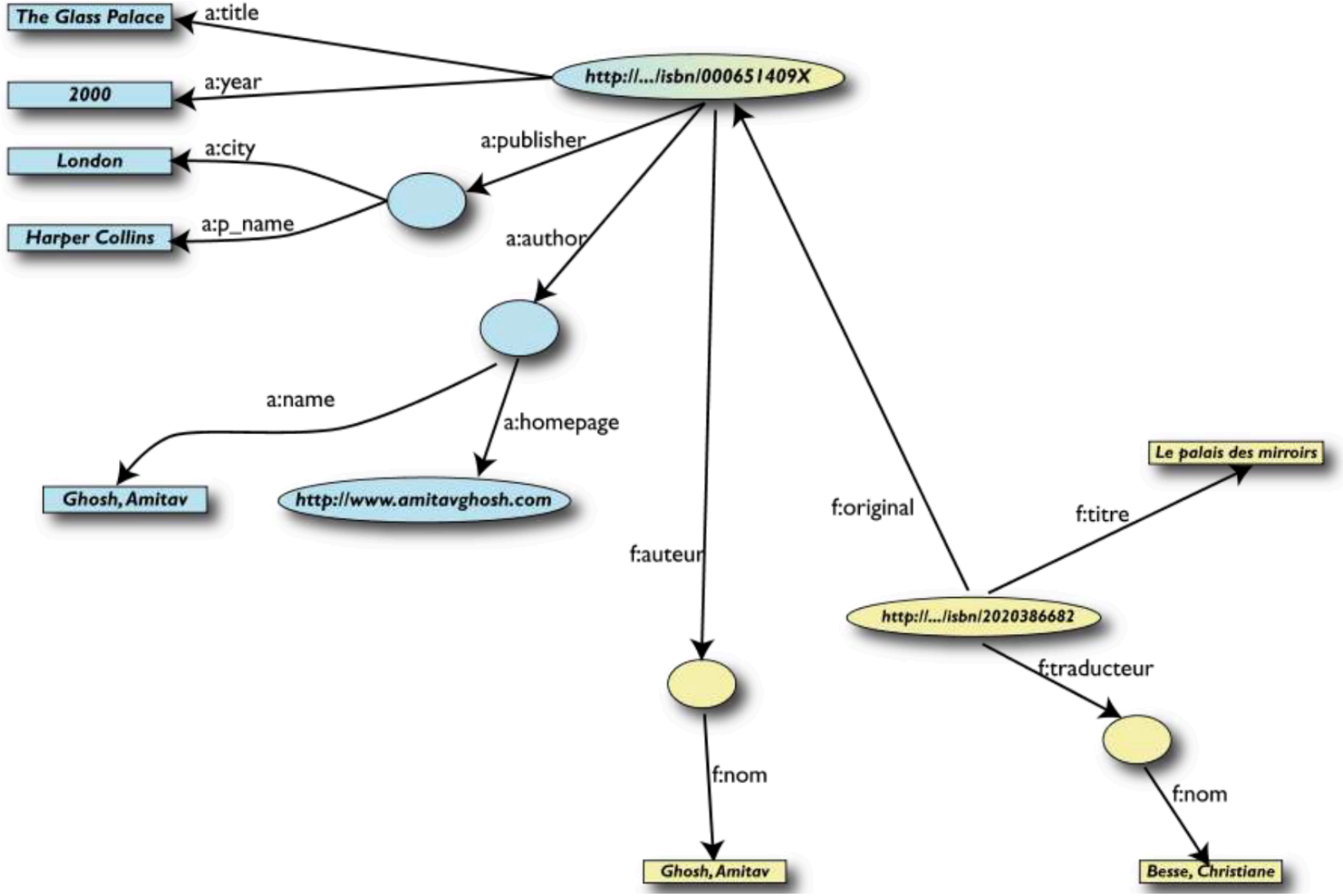
3rd: start merging data



3rd: start merging data



4th: merge identical resources

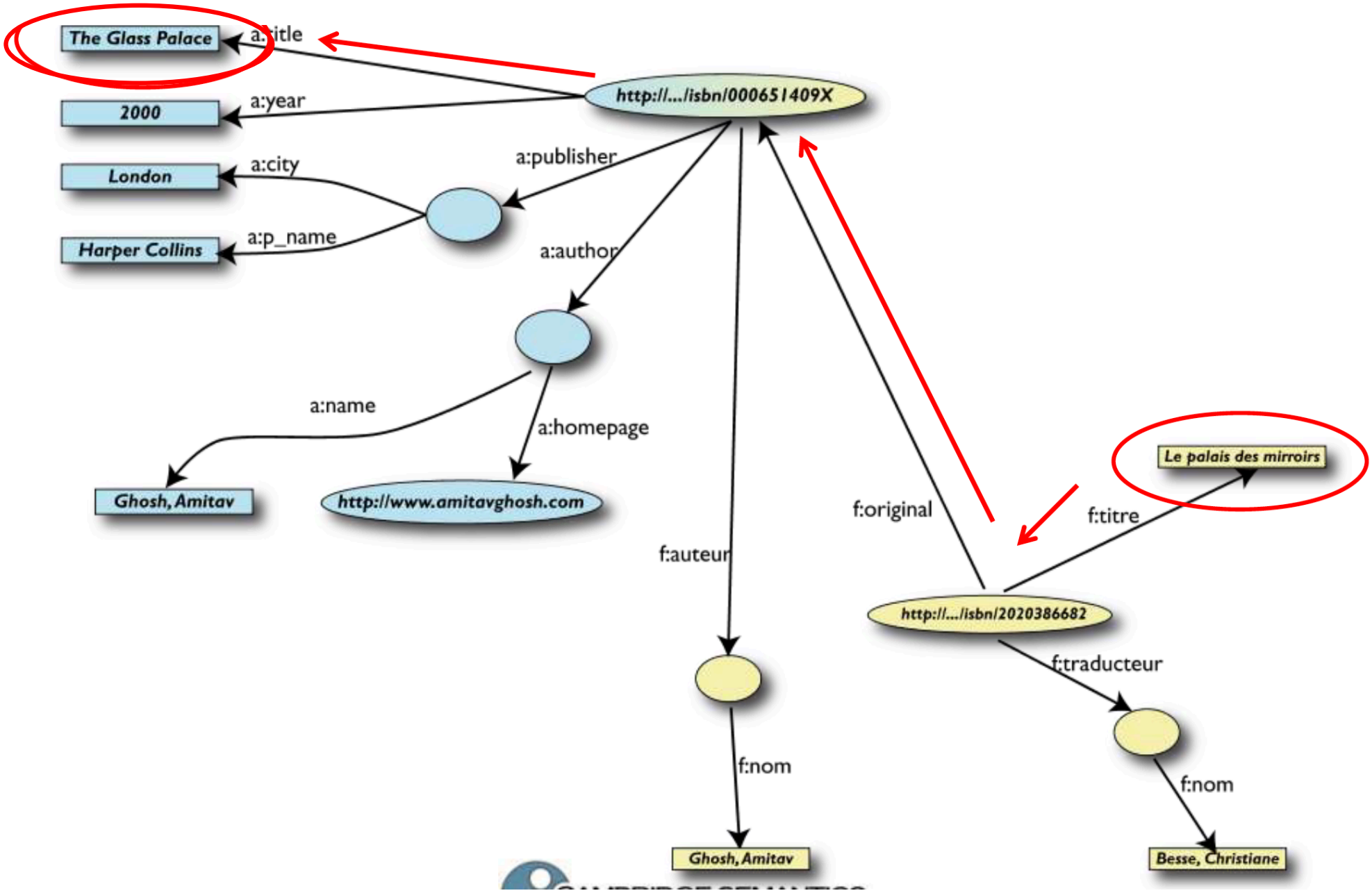


Start executing queries …

- User of data set “F” can now ask queries like:
  - “What is the title of the original version of Le “Palais des miroirs?”
- This information is not in the data set “F”…

…but can be retrieved after merging with data set “A”!

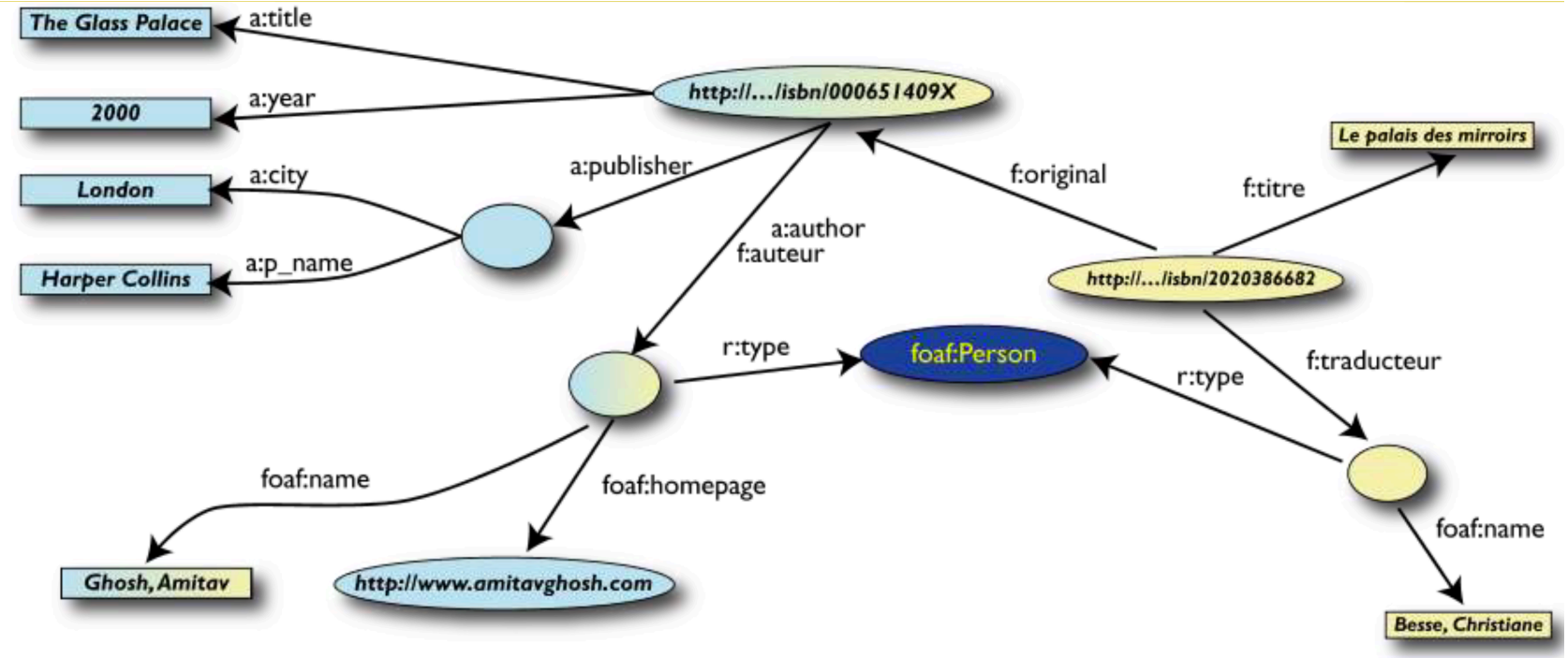
5th: query the merged dataset

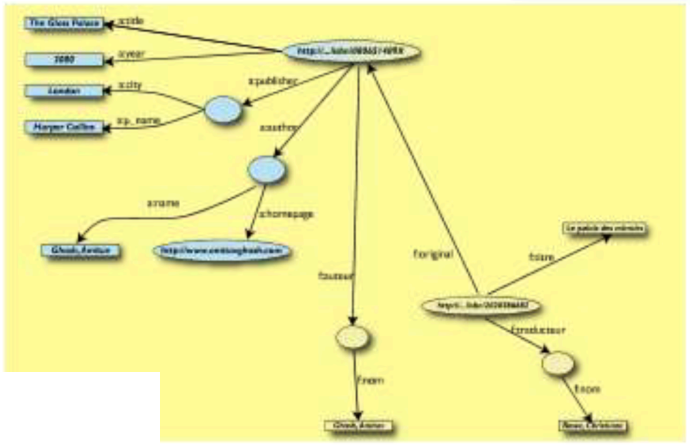


More can be achieved …

- We “know” that a:author and f:auteur are really the same
- But our automatic merge does not know that!
- Let us add some extra information to the merged data:
  - a:author is the same as f:auteur
  - Both identify a Person, a category (type) for certain resources
  - Person is an ontology concept

3rd revisited: use the extra knowledge (add ontology schema information)

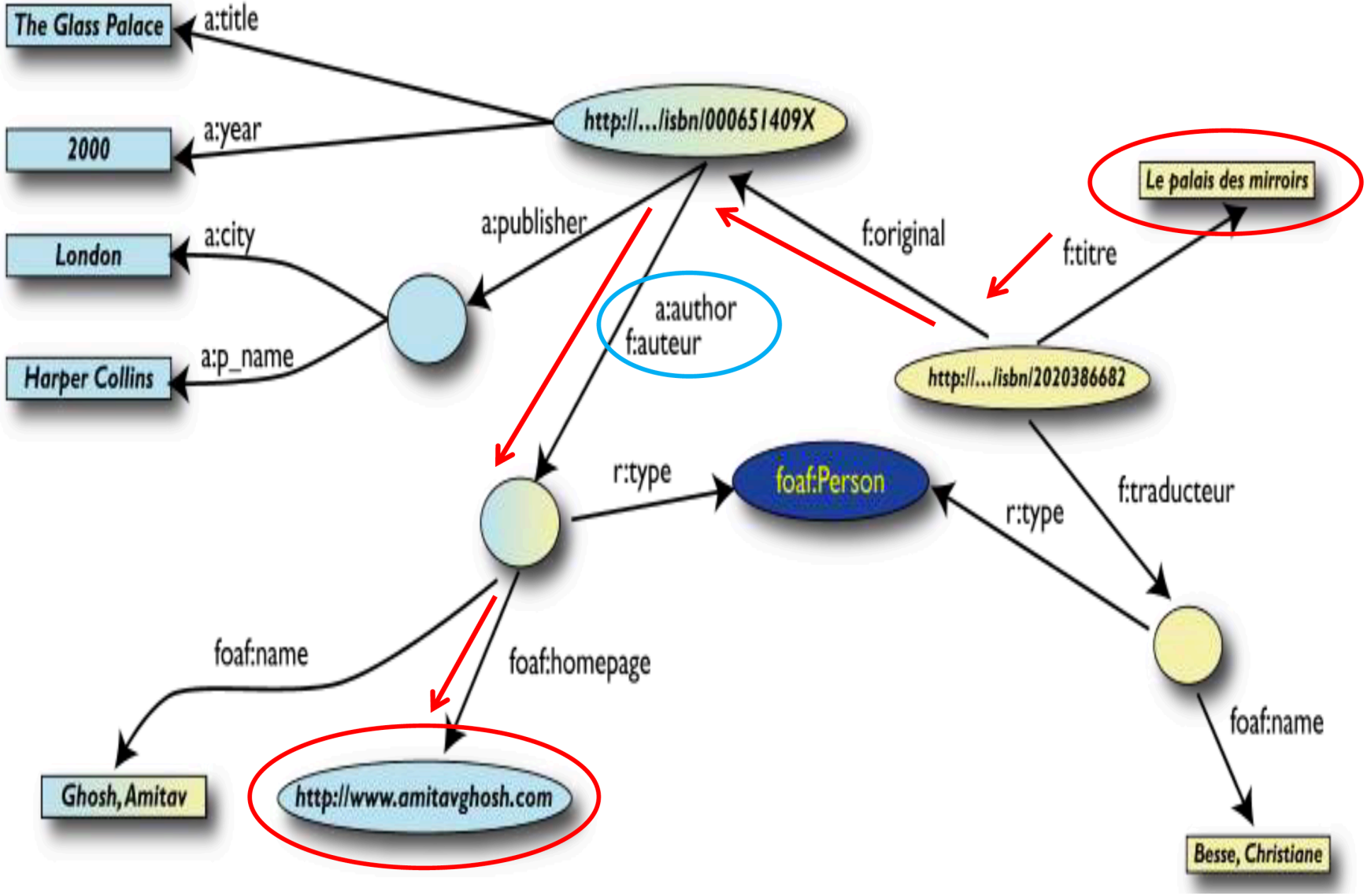




Start making richer queries

- User of data set “F” can now query:
  - “What is the home page of Le Palais des miroirs’s ‘auteur’?”
- The information is not in data set “F” or “A”…
- …but was made available by:
  - Merging data sets “A” and “F”
  - Adding three simple “glue” statements

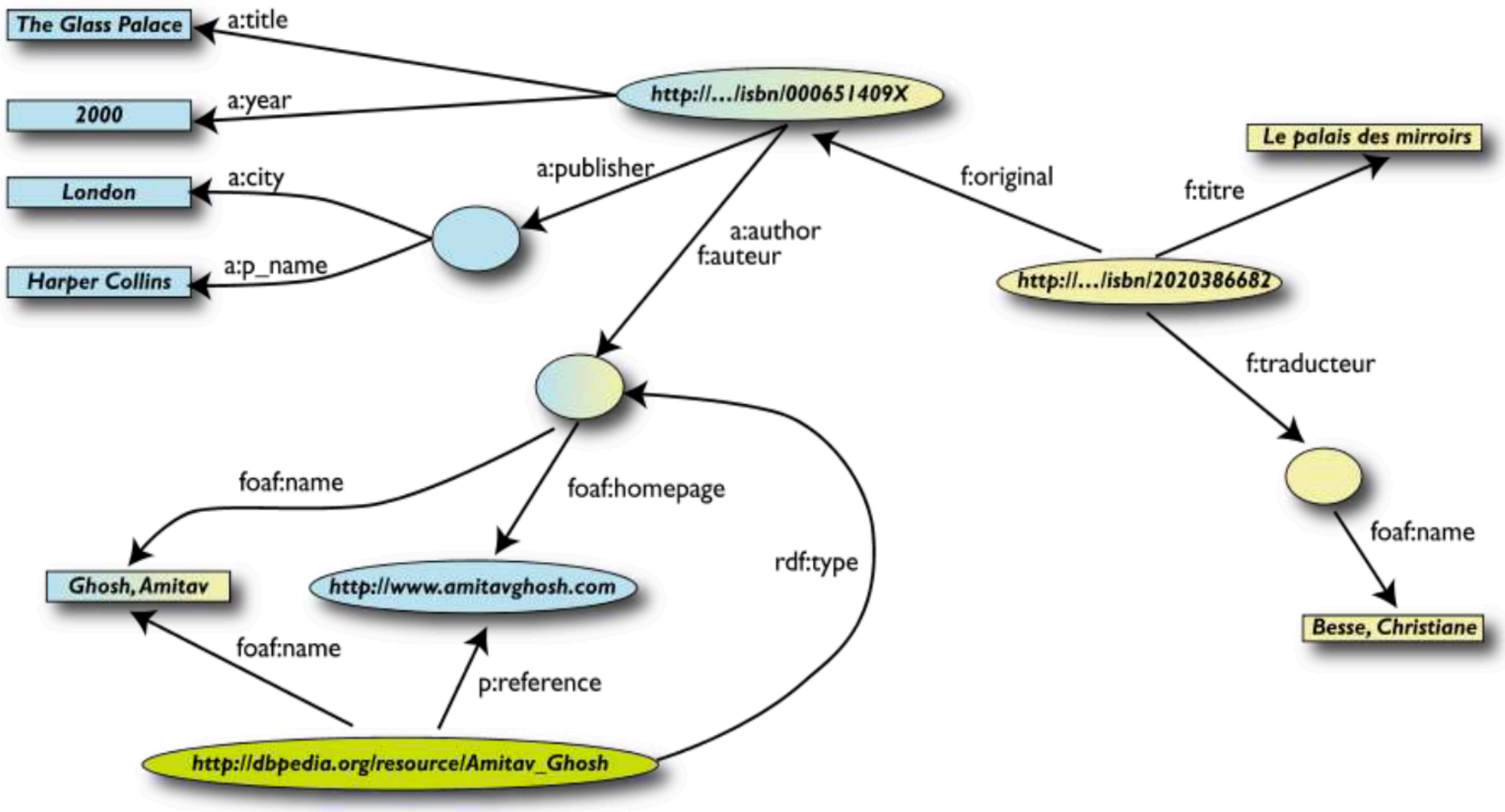
Richer queries



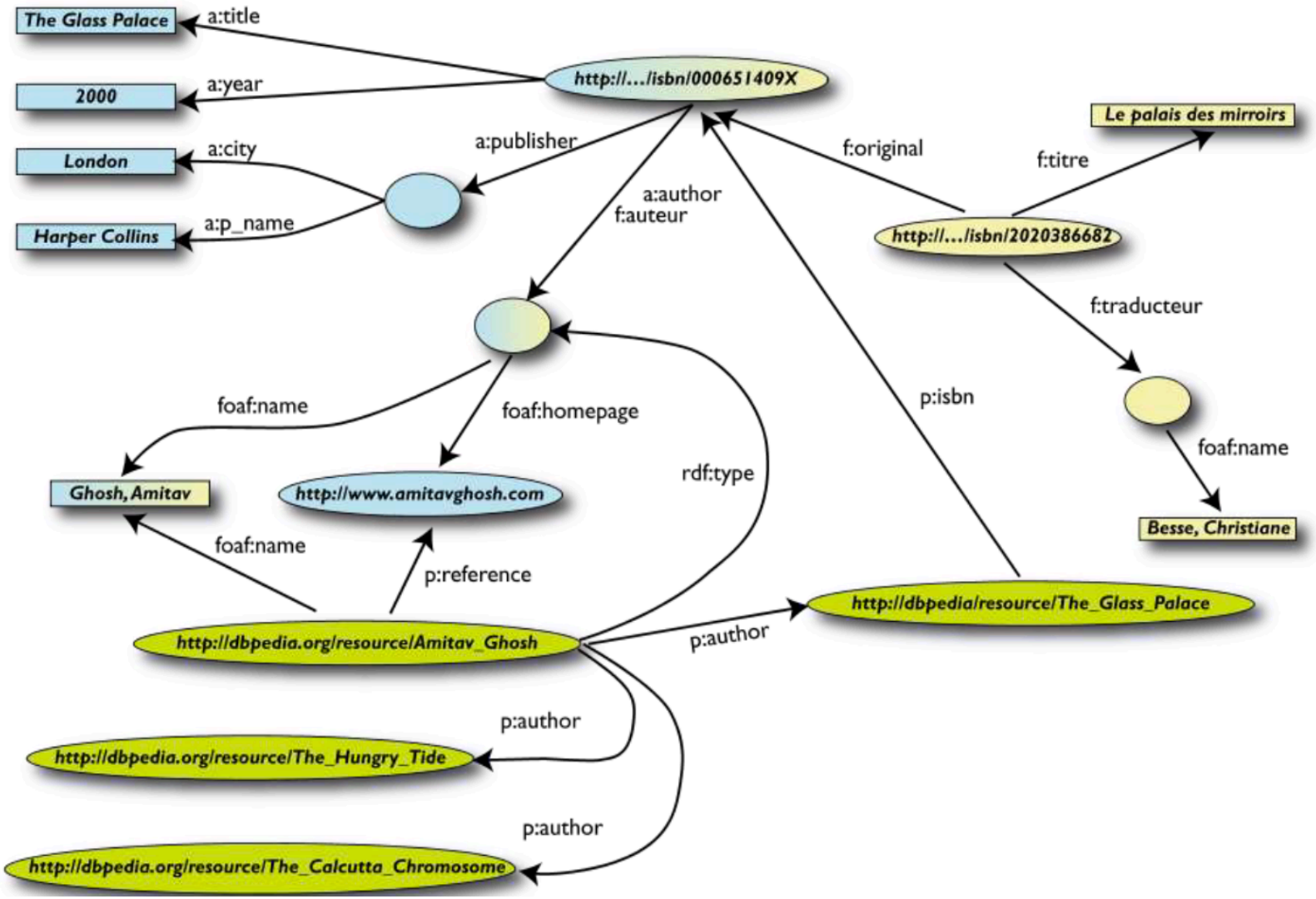
Bring in other data sources

- We can integrate new information into our merged data set from other sources
  - e.g. additional information about author Amitav Ghosh
- Perhaps the largest public source of general knowledge is Wikipedia
- Structured data can be extracted from Wikipedia using dedicated tools (leading to the dbpedia data source)

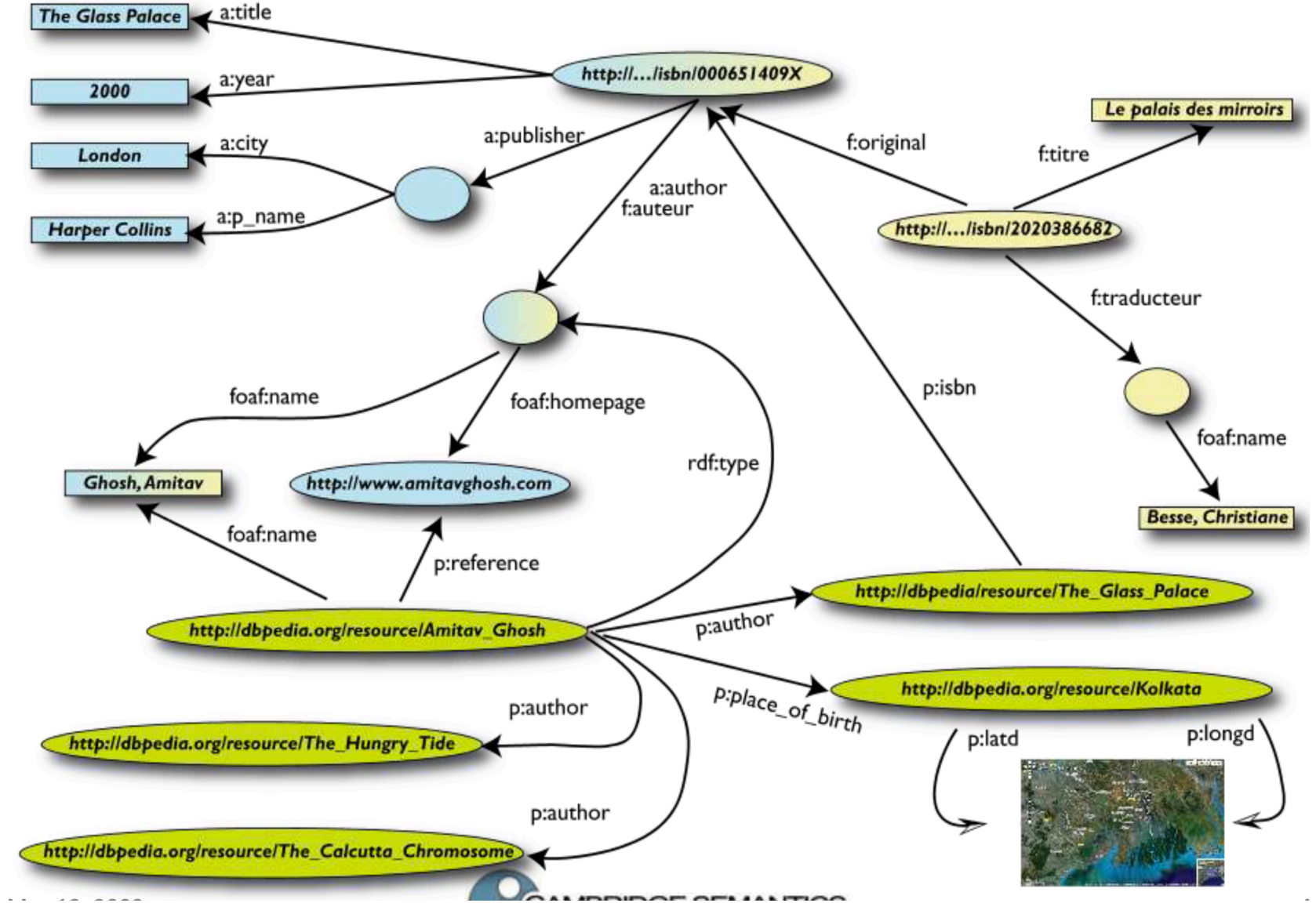
7th: merge with Wikipedia data



7th: merge with Wikipedia data



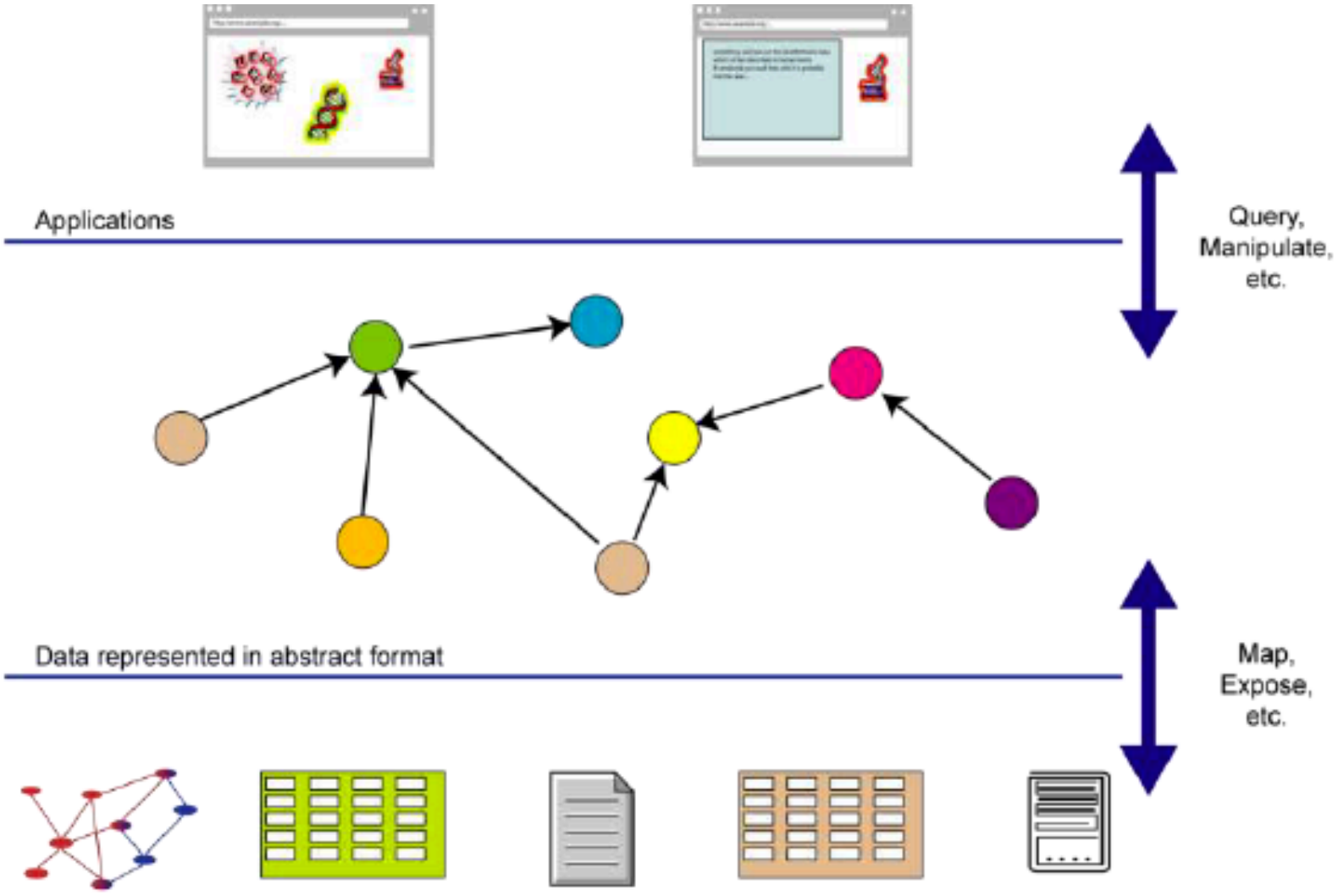
7th: merge with Wikipedia data



What did we do?

- We combined different data sets that
  - ...may be internal or somewhere on the Web
  - ...are of different formats (RDBMS, Excel spreadsheet, (X)HTML, etc)
  - ...have different names for the same relations
- We could combine the data because some identifiers were identical i.e. the ISBNs in this case (i.e., they correspond to the same ontology instance)
- We could add some simple additional information (the “glue”) to help further merge data sets (through new ontology instances, ontology relations, ontology concepts)
- The result? Answer queries that could not previously be asked

What did we do?



The abstraction pays off because …

- …the graph representation is independent of the details of the native structures
- …a change in local database schemas, HTML structures, etc. do not affect the whole "schema independence"
- …new data, new connections can be added seamlessly & incrementally

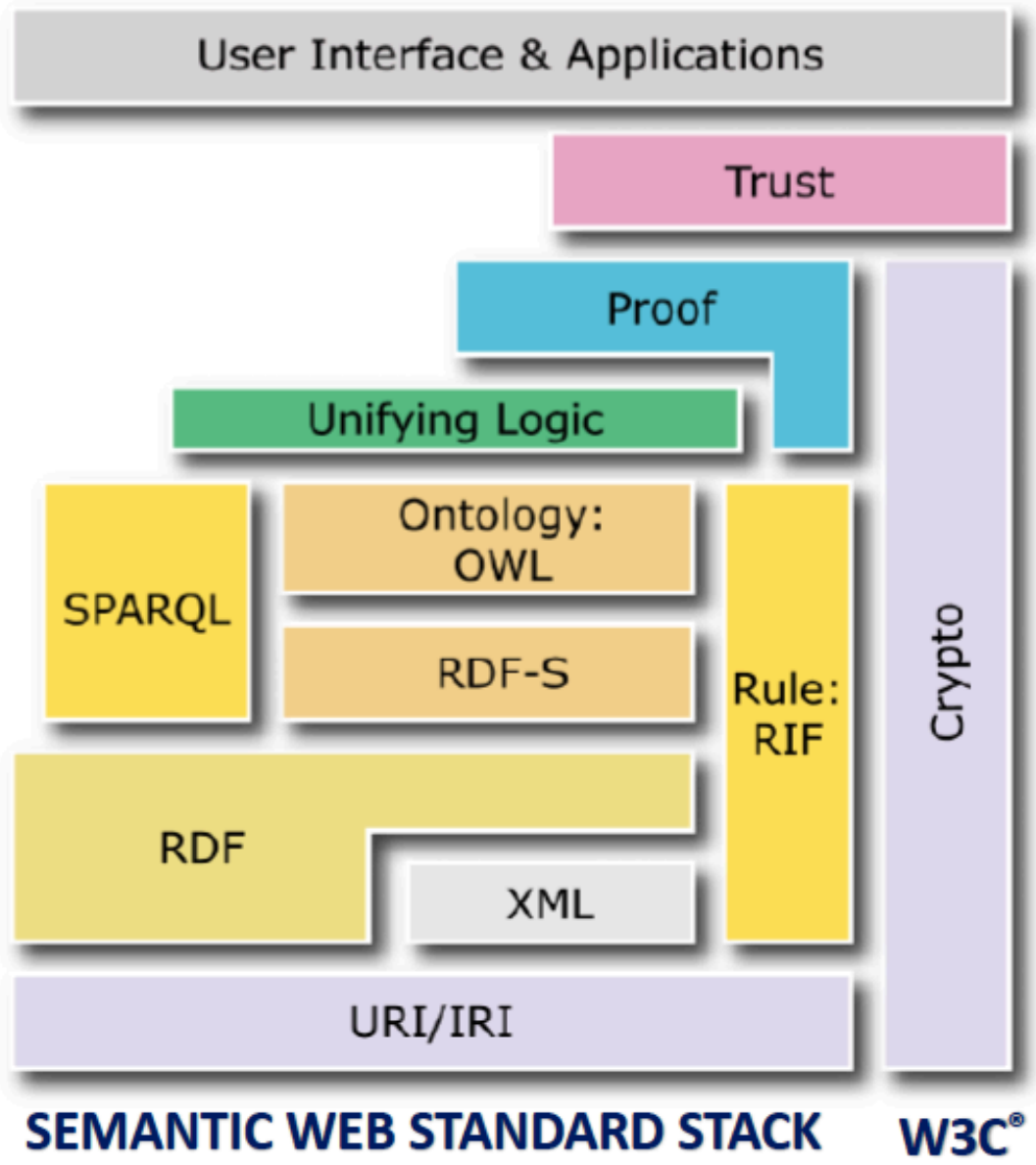
Semantic Web technologies make such integration possible

The Semantic Web

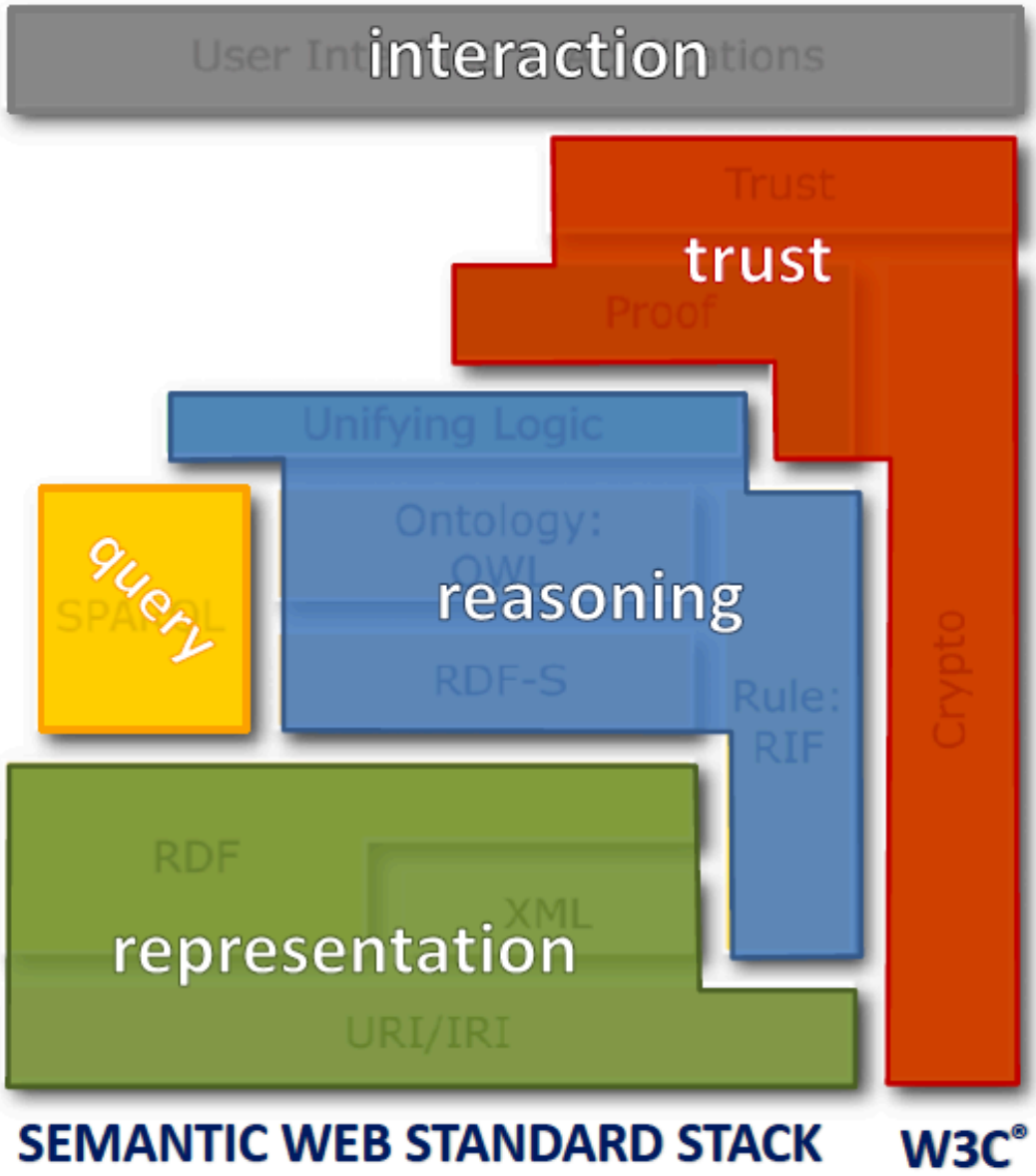
- The Semantic Web is a Web in which the resources (things) are semantically described, through the usage of an ontology
- A resource is anything that can be referred to by a URI (Uniform Resource Identifiers)
  - a web page, identified by an URL
  - a fragment of an XML document, identified by an element node of the document
  - a web service
  - a thing, an object, a property, etc.
- Examples
  - http://www.example.org/file.html#home
  - http://www.example.org/file2.xml#xpath(//q[@a=b])
  - http://www.example.org/form?a=b&c=d

"Now! *That* should clear up a few things around here!"

The Semantic Web Standard Stack



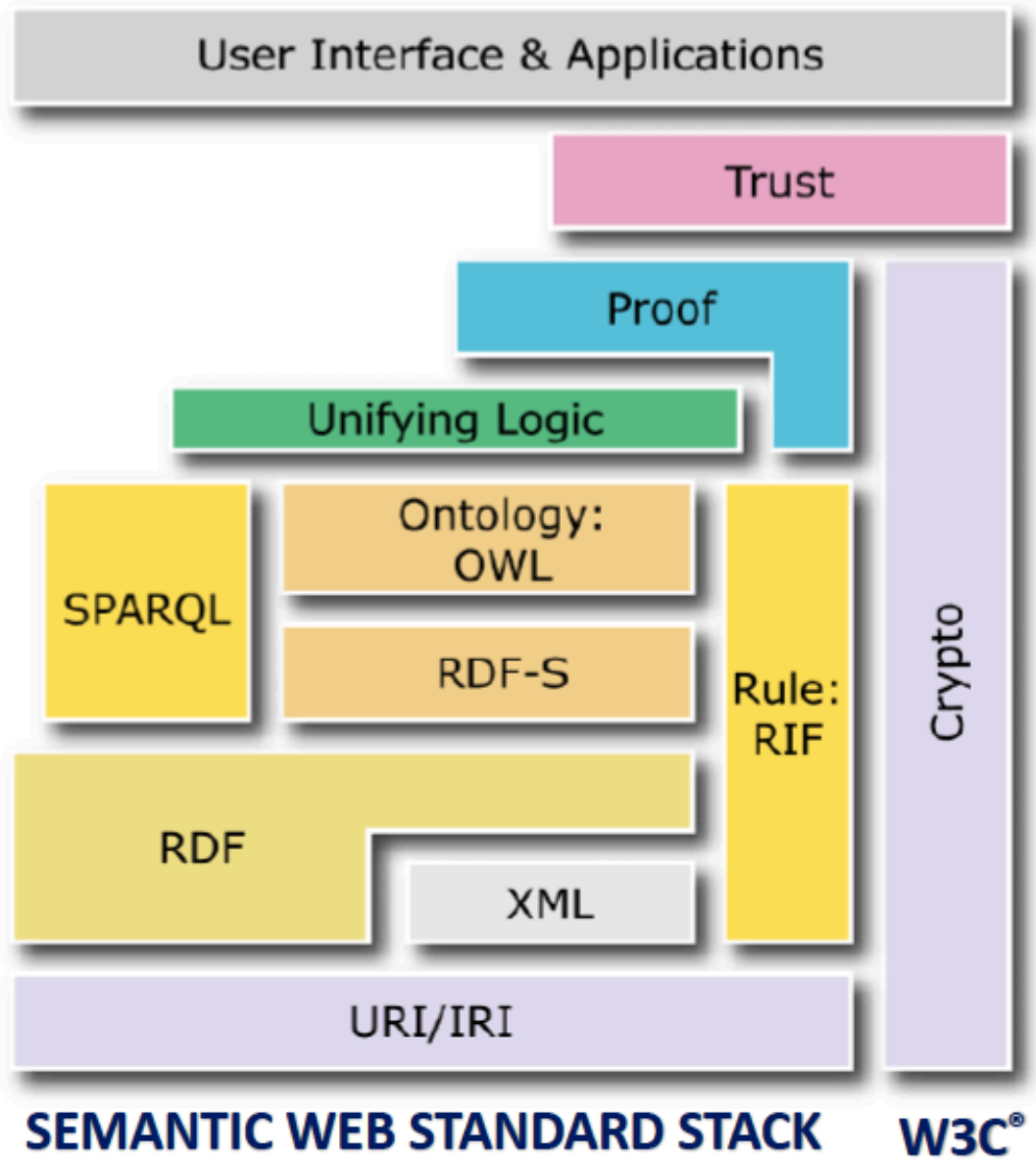
The Semantic Web Standard Stack



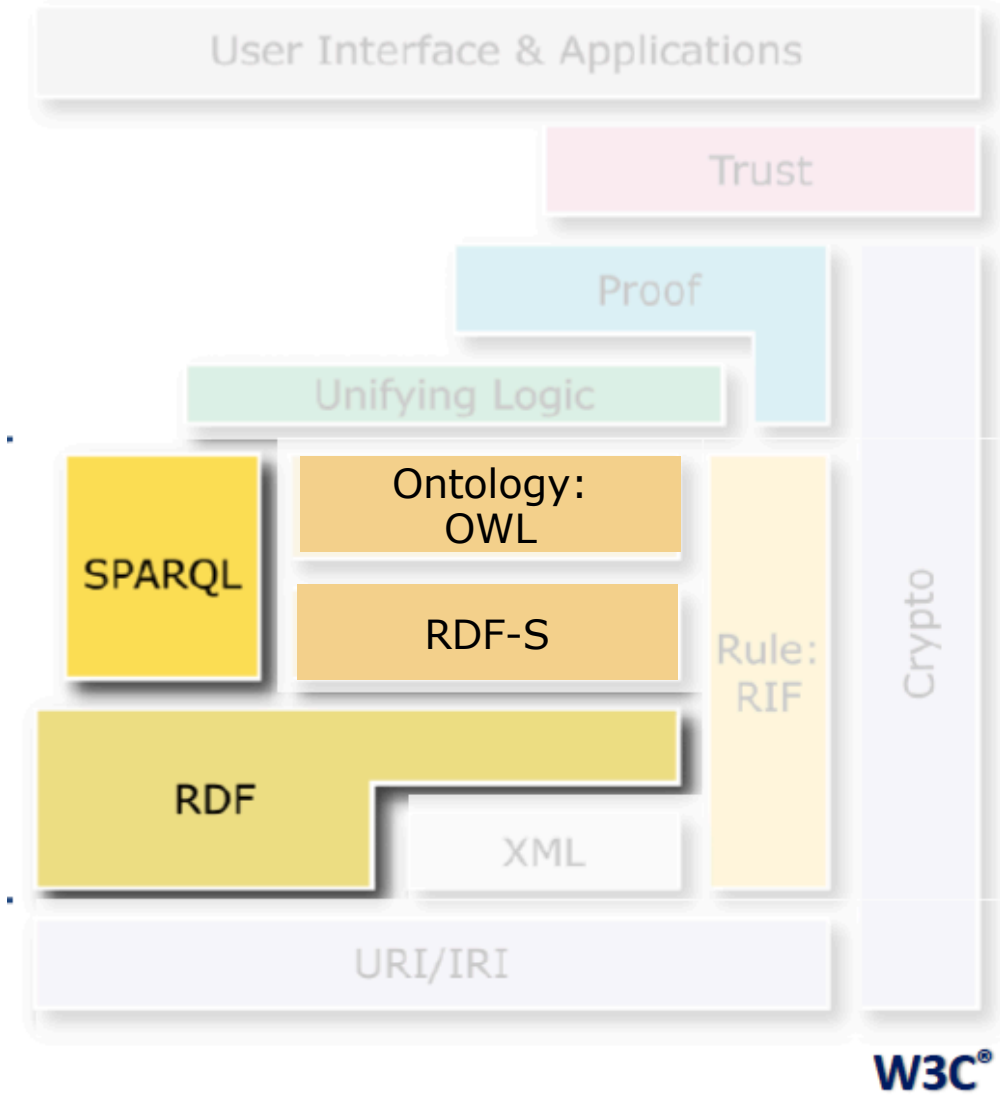
Ontology Languages for the Semantic Web

- RDF: a very simple ontology language
- RDFS: Schema for RDF
  - Can be used to define richer ontologies
- OWL: a much richer ontology language

The Semantic Web Standard Stack



The Semantic Web Standard Stack



W3C Linked Open Data Project

- Goal: “expose” open datasets in RDF on the Web
- Set RDF links among the data items from different datasets
- Set up, if possible, query endpoints

Linked Open Data

- **Linked data** is just RDF (RDFS, OWL) data, linking together resources potentially described in different datasets
- **Linked open** data is just linked data freely accessible on the Web along with any required ontologies

Properties of Linked (Open) Data

- Anyone can publish data on the Web of Linked Data
- Entities are connected by links
  - Creating a global data graph that spans data sources and enables the discovery of new data sources.
- Data is self-describing
  - If an application encounters data represented using an unfamiliar vocabulary, the application can resolve the URIs that identify vocabulary terms in order to find their RDFS or OWL definition.
- The Web of Data is open
  - Meaning that applications can discover new data sources at run-time by following links.

Example: DBpedia

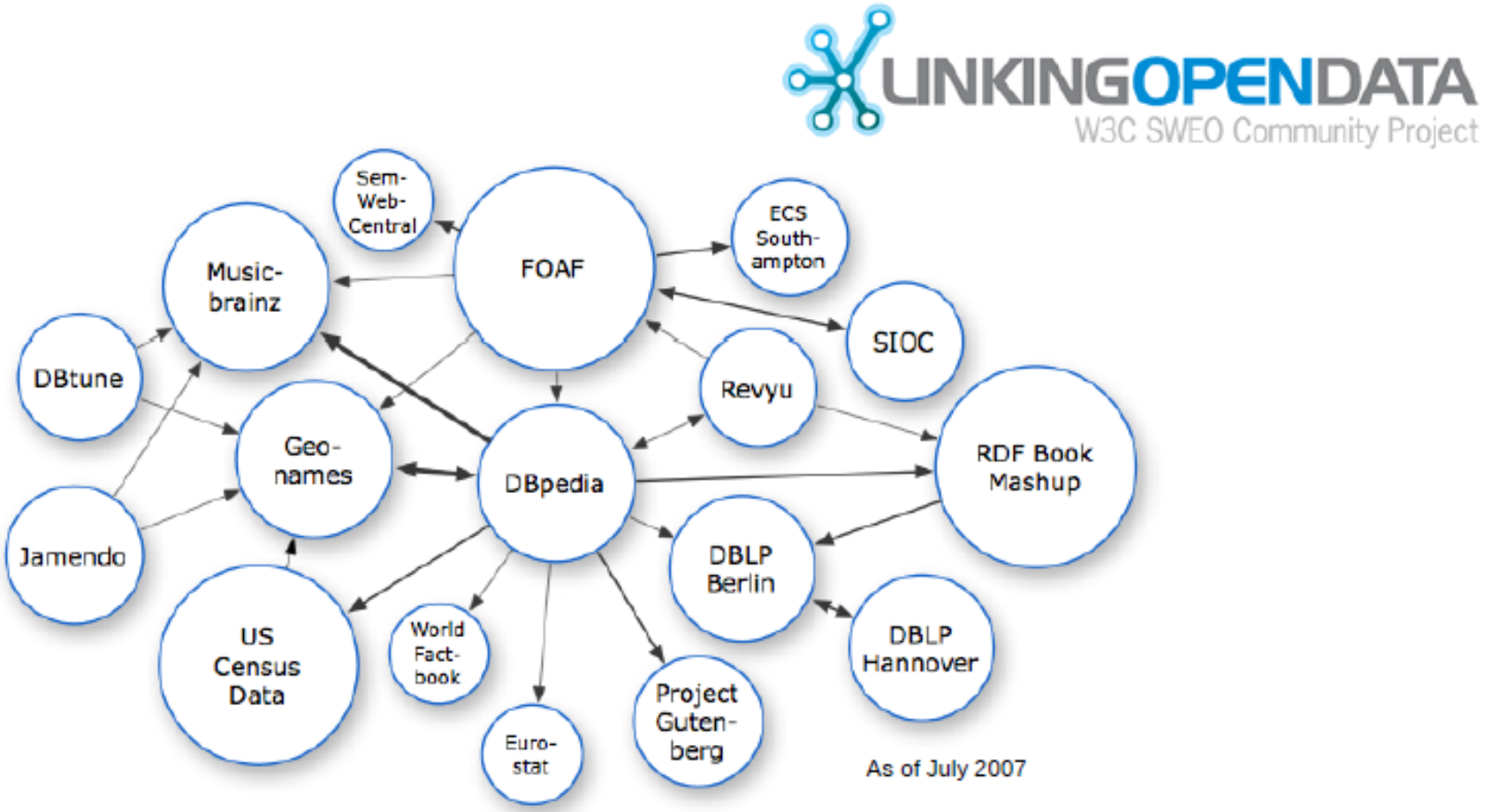
- DBpedia is a community effort to
- extract structured (“infobox”) information from Wikipedia
- provide a query endpoint to the dataset
- interlink the DBpedia dataset with other datasets on the Web

Example: DBPedia

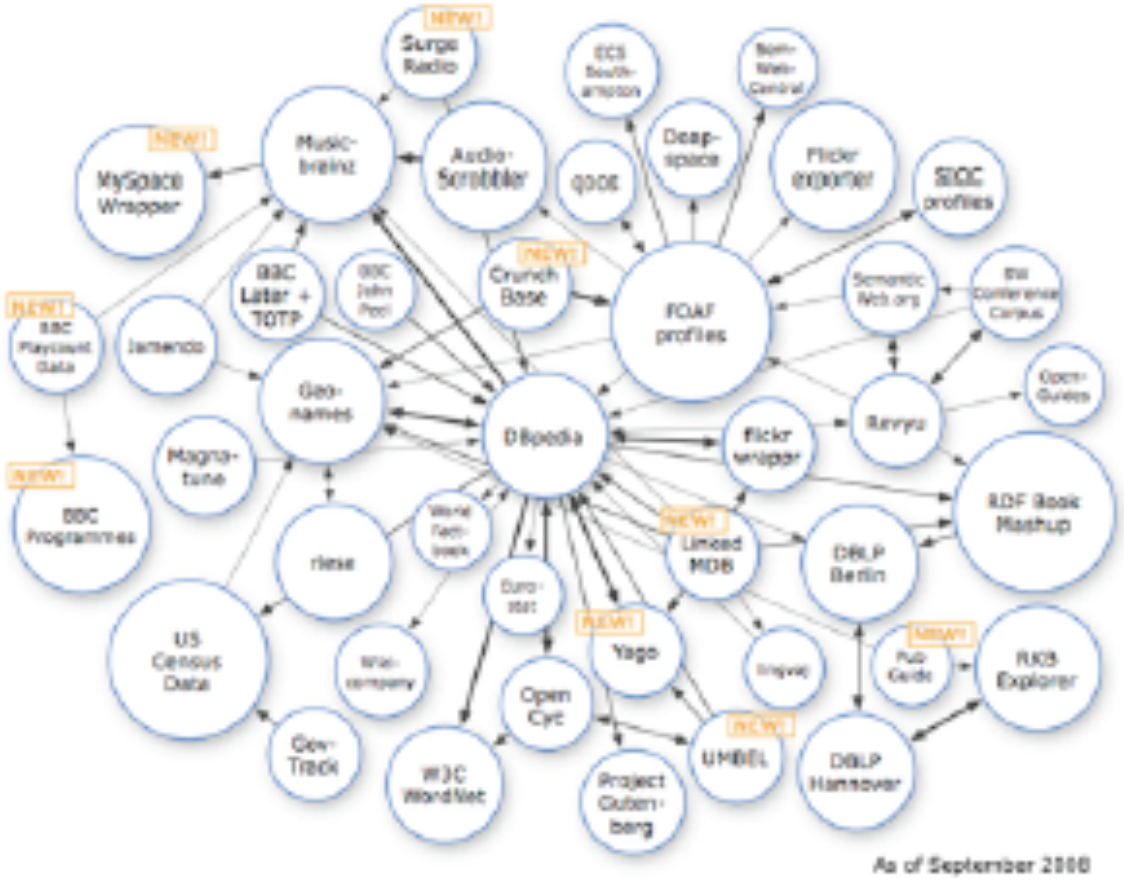
```
@prefix dbpedia <http://dbpedia.org/resource/>.
@prefix dbterm  <http://dbpedia.org/property/>.

dbpedia:Amsterdam
  dbterm:officialName "Amsterdam" ;
  dbterm:longd "4" ;
  dbterm:longm "53" ;
  dbterm:longs "32" ;
  dbterm:leaderName dbpedia:Lodewijk_Asscher ;
  ...
  dbterm:areaTotalKm "219" ;
  ...
dbpedia:ABN_AMRO
  dbterm:location dbpedia:Amsterdam ;
  ...
```

Linked Open Data Project

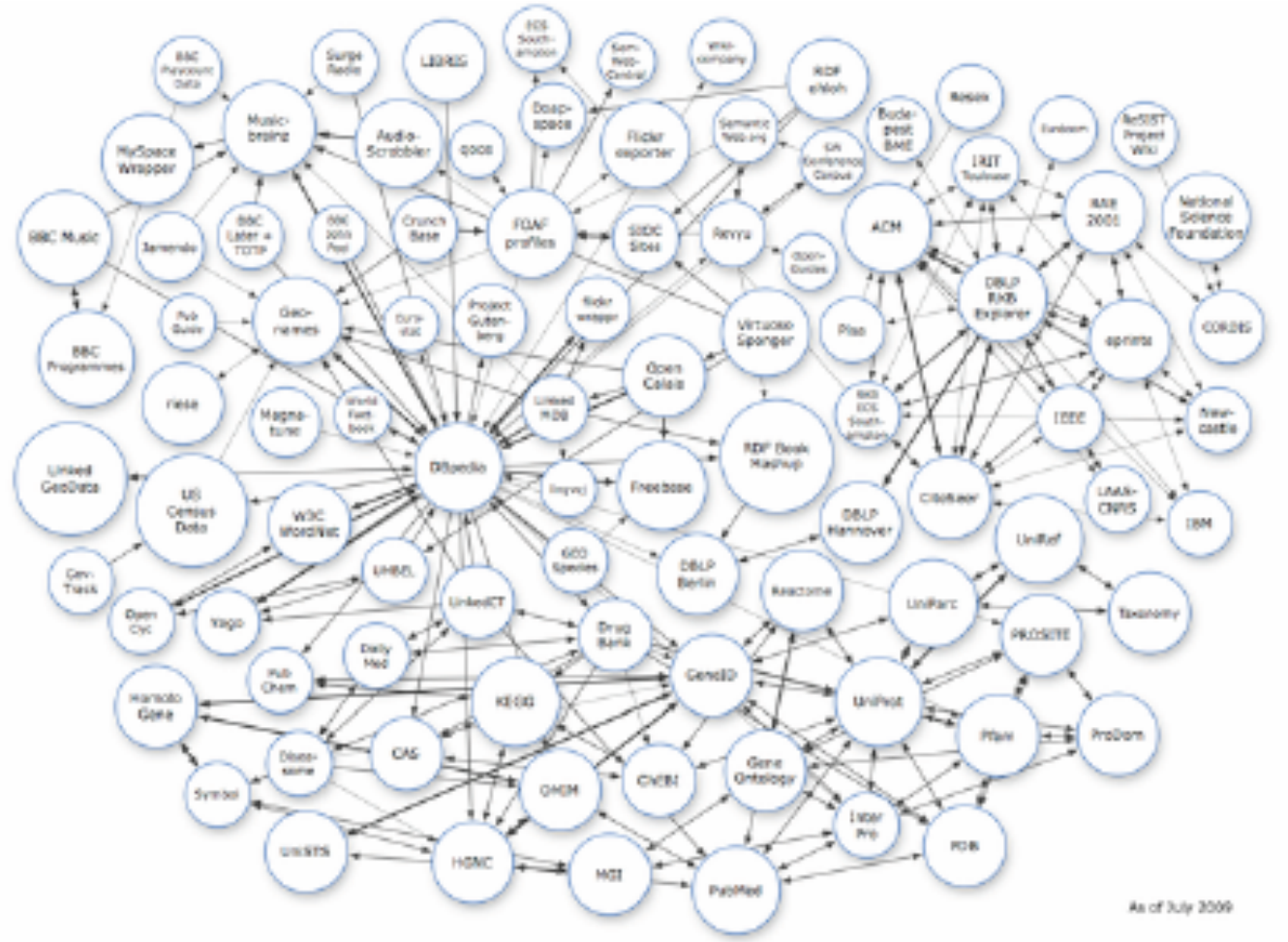


Linked Open Data Project



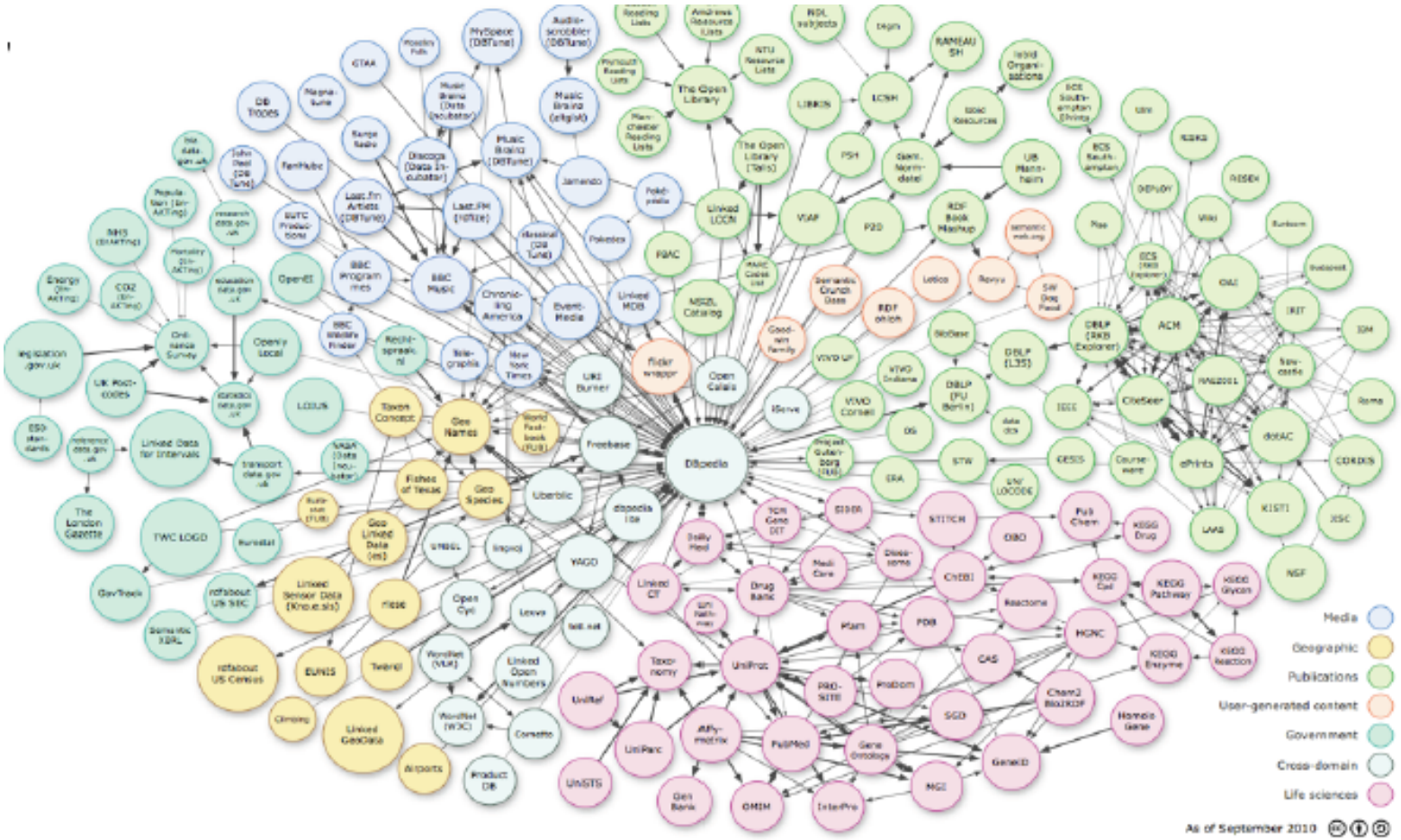
2008

Linked Open Data Project



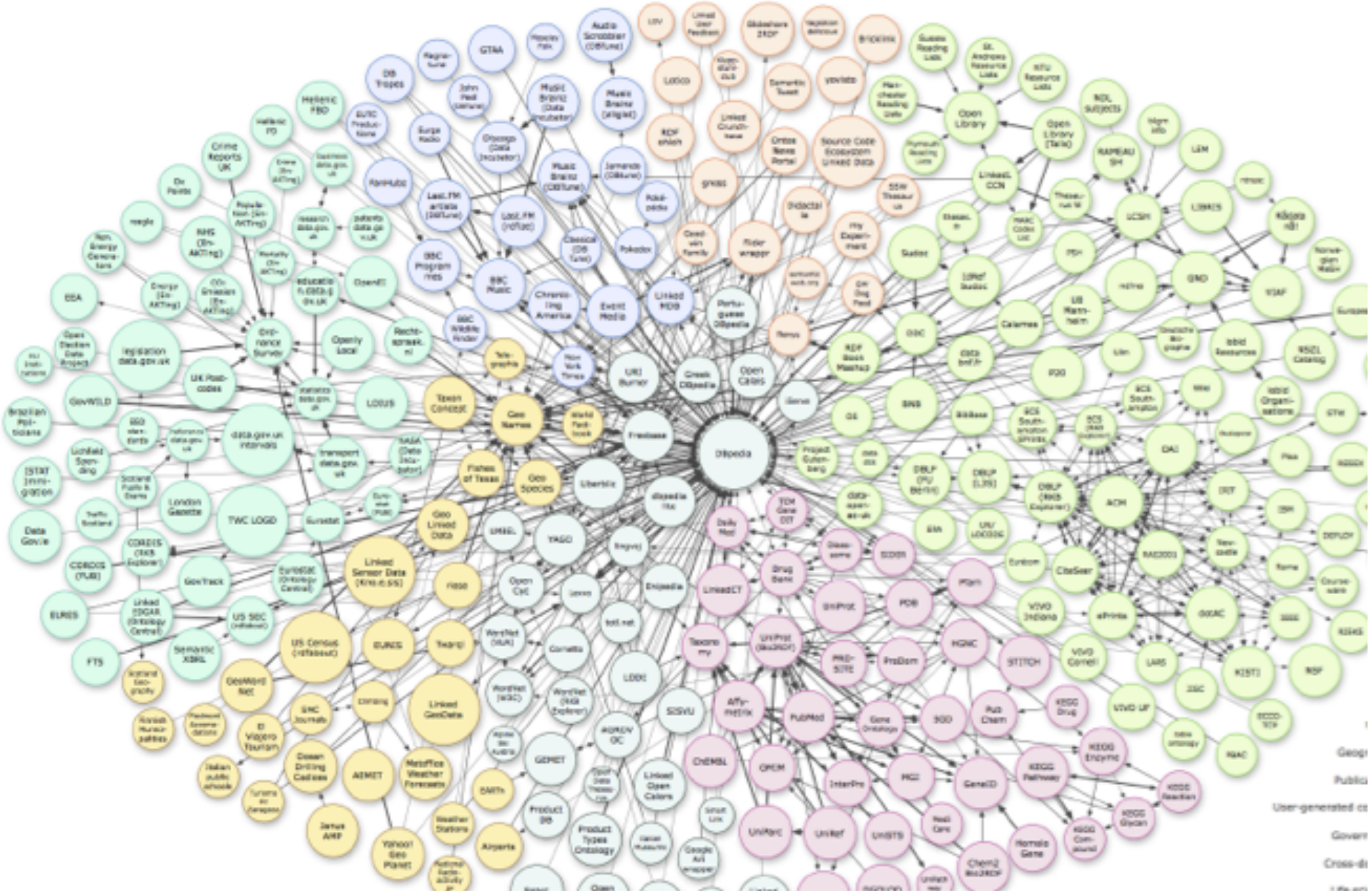
2009

Linked Open Data Project

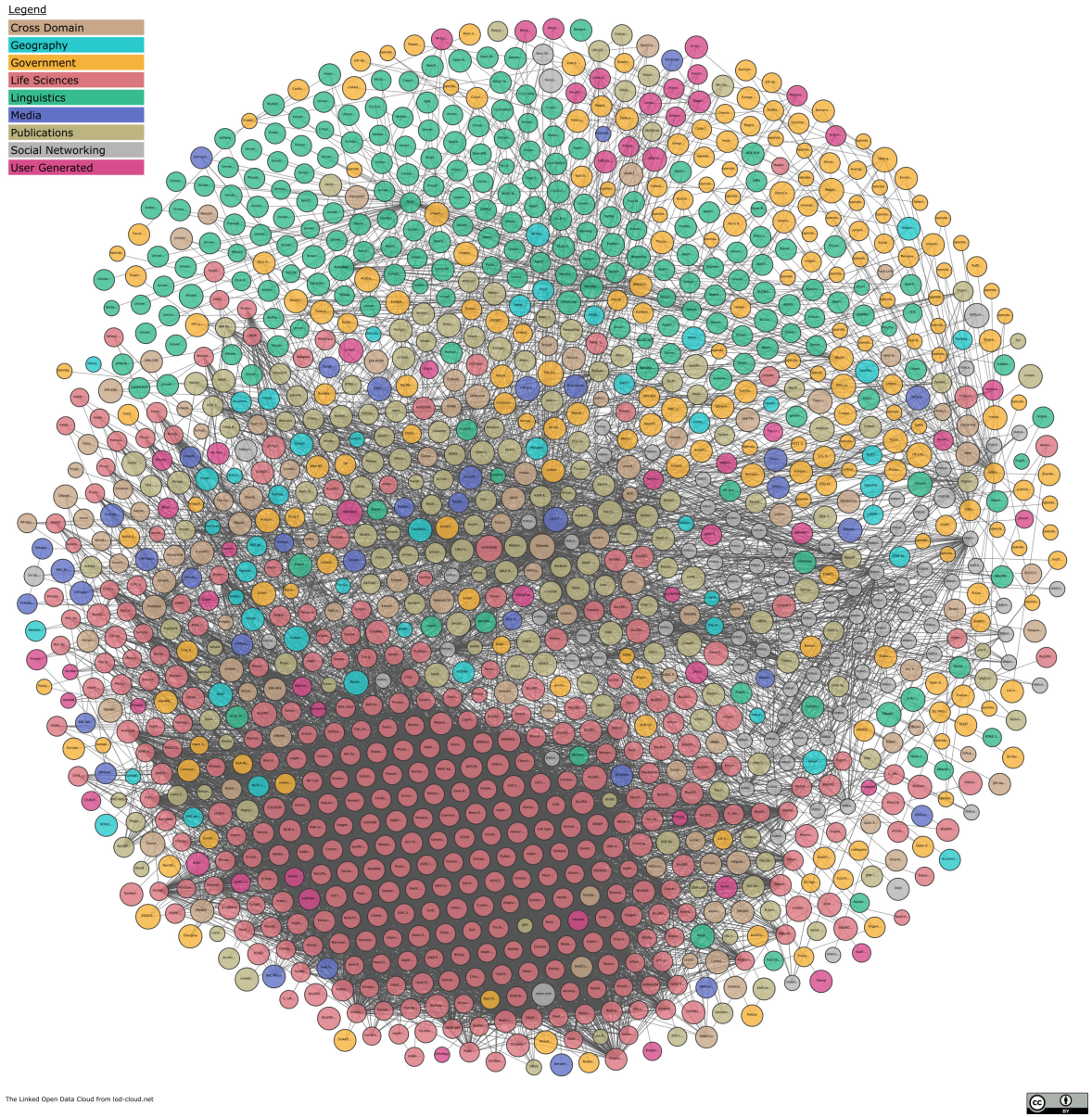


2010

Linked Open Data Project



Linked Open Data Project



The **dataset** contains 1269 datasets with 16201 links

(as of May 2020)

https://lod-cloud.net