Implications for the future

Dave Raggett, W3C

dsr@w3.org

NSF Workshop on Knowledge Representation
and Information Management for Financial Risk
Management, 21-22 July, Arlington, VA

http://irix.umiacs.umd.edu/

21 July 2010

With thanks to Ivan Herman for use of some of his slides

W3C
Semantic Web

Before and after – impact of the Web

- Before the Web, applications were all locally installed and operated off local data

- The Internet and the Web changed all that
  - Browser as ubiquitous client
  - HTTP as the interface to remote services
  - Server-side scripts for access to legacy systems
    - Hiding the internal data formats and interfaces
  - Explosion of entrepreneurial activity
    - Due to ease in reaching potential customers
  - Rise of cloud computing and software as service

W3C Semantic Web

The Unfinished Revolution

- Today's Web is designed for people to interpret

  - Using your eyes and your mind

- Each website only covers part of your needs

  - You have to do integrate information across websites

  - This is time consuming and a waste of effort

- We should put computers to work on our behalf

  - We need to find ways for software to query, combine and interpret data accessible over the Web

    - Michael Dertouzos: *"The Unfinished Revolution, How to Make Technology Work for Us--Instead of the Other Way Around"*

Implications for Financial Risk Management

- The Web makes it easier to create applications acting over information
  - from different parts of an organization
  - and across multiple organizations
- Building upon, rather than replacing existing systems
  - HTTP with scripts and query languages as transducers
  - Exploiting investment in existing relational databases
- Greater transparency of operations
  - Potential for improved utilization of financial resources
- But dependent on freedom to innovate
  - Evolution versus intelligent design

W3C **Semantic Web**

W3C/XBRL Int. inc Workshop on improving access to financial information on the Web

5-6 October 2009, Arlington VA, hosted by the FDIC

http://www.w3.org/2009/03/xbrl/cfp

W3C/XII 2009 Workshop

- Focused on opportunities and challenges for interactive access to business and financial data

- Brought together people from a wide range of backgrounds
  - Including government agencies (SEC, FDIC, FRB, EPA, FSTC, NIEM), businesses and academic researchers across the World
  - Sharing experiences in XBRL, Semantic Web and other fields

- Identified challenges

  - Practices for naming business and financial entities and associated metadata as a basis for comparing and combining different sources of information
  - Practices for harmonizing vocabularies, and the need for a continuing dialog across government agencies and business organizations
  - Need for robust treatment of provenance to avoid abuses
  - Further technical work on extending OWL to support richer integrity constraints, and role of intermediate data models for simplifying application development

More details at   http://www.w3.org/2009/03/xbrl/report

So what is the Semantic Web?

It is, essentially, the Web of Data and the technologies to realize that

Is it that simple...

- Of course, the devil is in the details
  - a common model has to be provided for machines to describe and query the data and its connections
  - the "classification" of the terms can become very complex for specific knowledge areas: this is where ontologies, thesauri, etc, enter the game…

Linked Data

Data Integration with the Semantic Web

- ## Map each data source into binary relations*
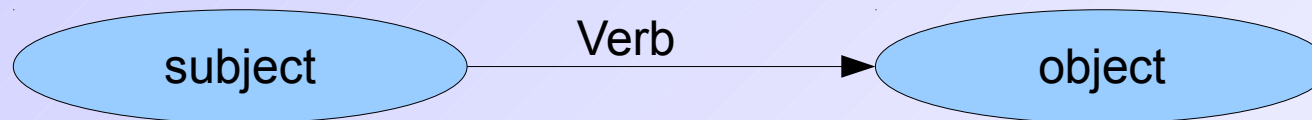
All three are named with URIs

- ## Merge the relations from different sources

- ## Start making queries

Binary relations as RDF triples
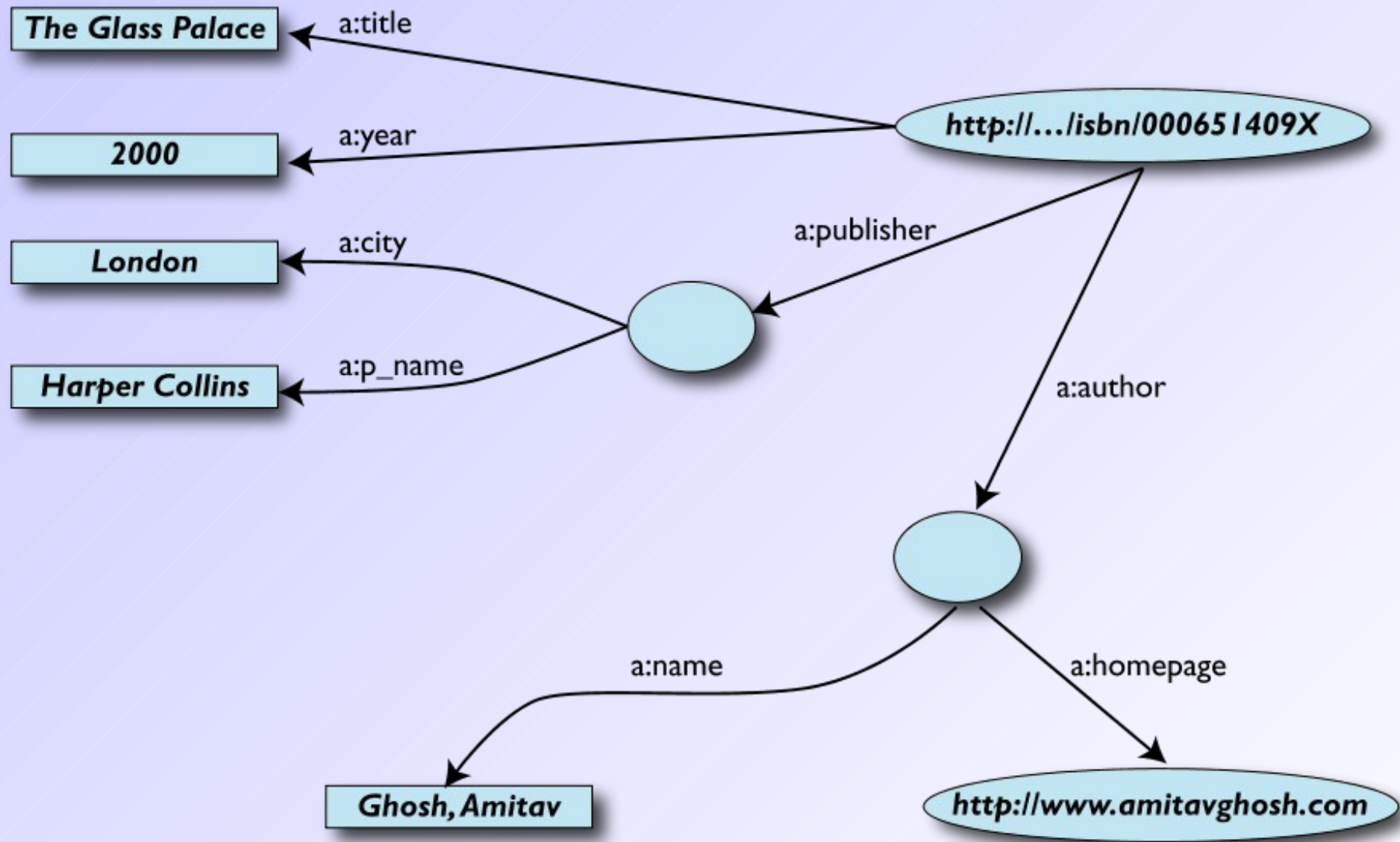
A simplified book store example

SQL database:

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	Harper Collins	London

Export data as relations

Another book store example

Spreadsheet

	A	B	D	E
1	**ID**	**Titre**	**Traducteur**	**Original**
2	ISBN0 2020386682	Le Palais des miroirs	A13	ISBN-0-00-651409-X
3				
6	**ID**	**Auteur**		
7	ISBN-0-00-651409-X	A12		
11	**Nom**			
12	Ghosh, Amitav			
13	Besse, Christianne			

Export it as relations
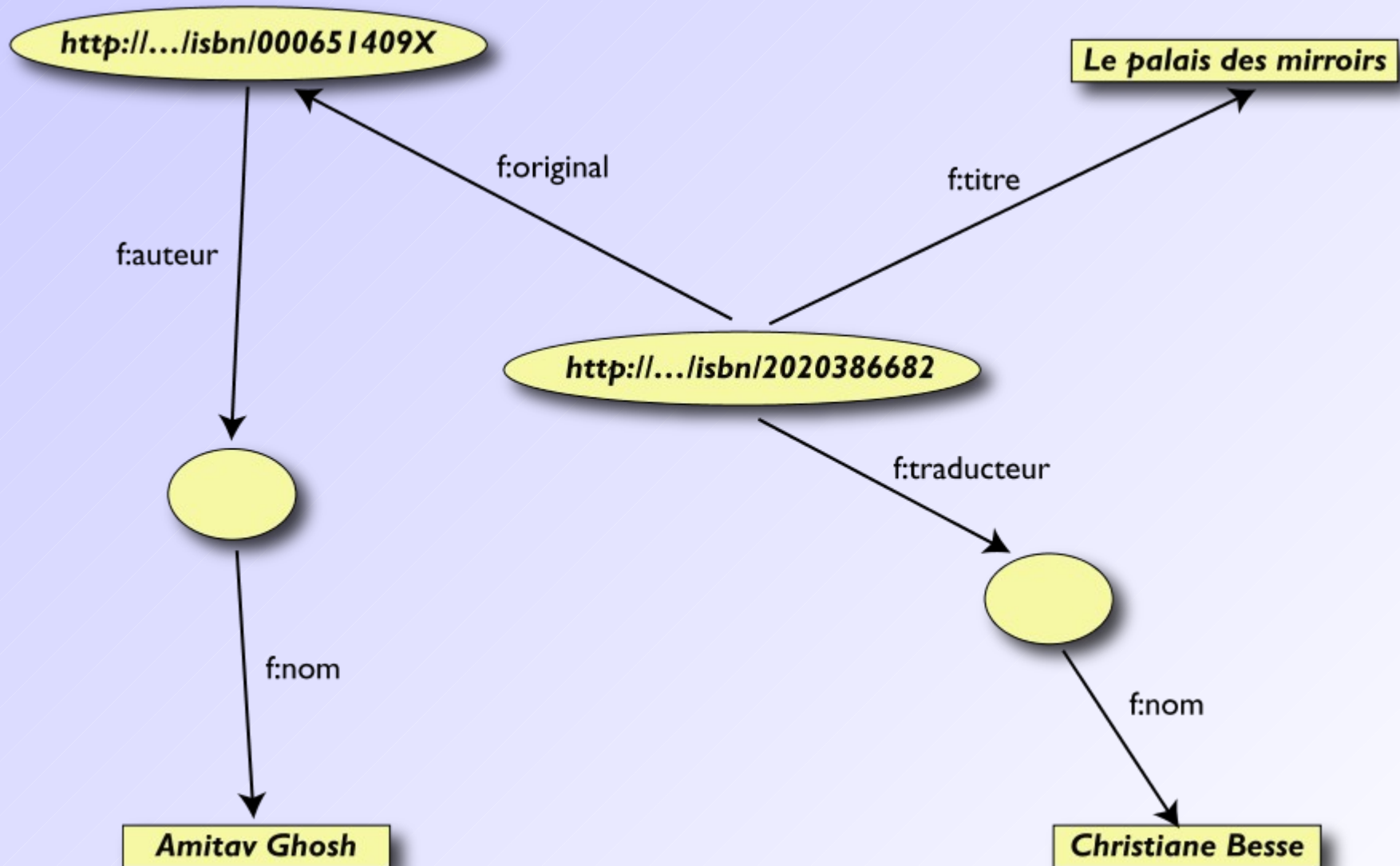
Merge the relations

Merging continued...

The Glass Palace — a:title

2000 — a:year

London — a:city

Harper Collins — a:p_name

http://...isbn/000651409X

a:publisher

a:author

a:name — Ghosh, Amitav

a:homepage — http://www.amitavghosh.com

Same URI = Same Resources

http://...isbn/000651409X

f:auteur

f:original

f:nom — Ghosh, Amitav

Le palais des mirroirs

f:titre

http://...isbn/2020386682

f:traducteur

f:nom — Besse, Christiane
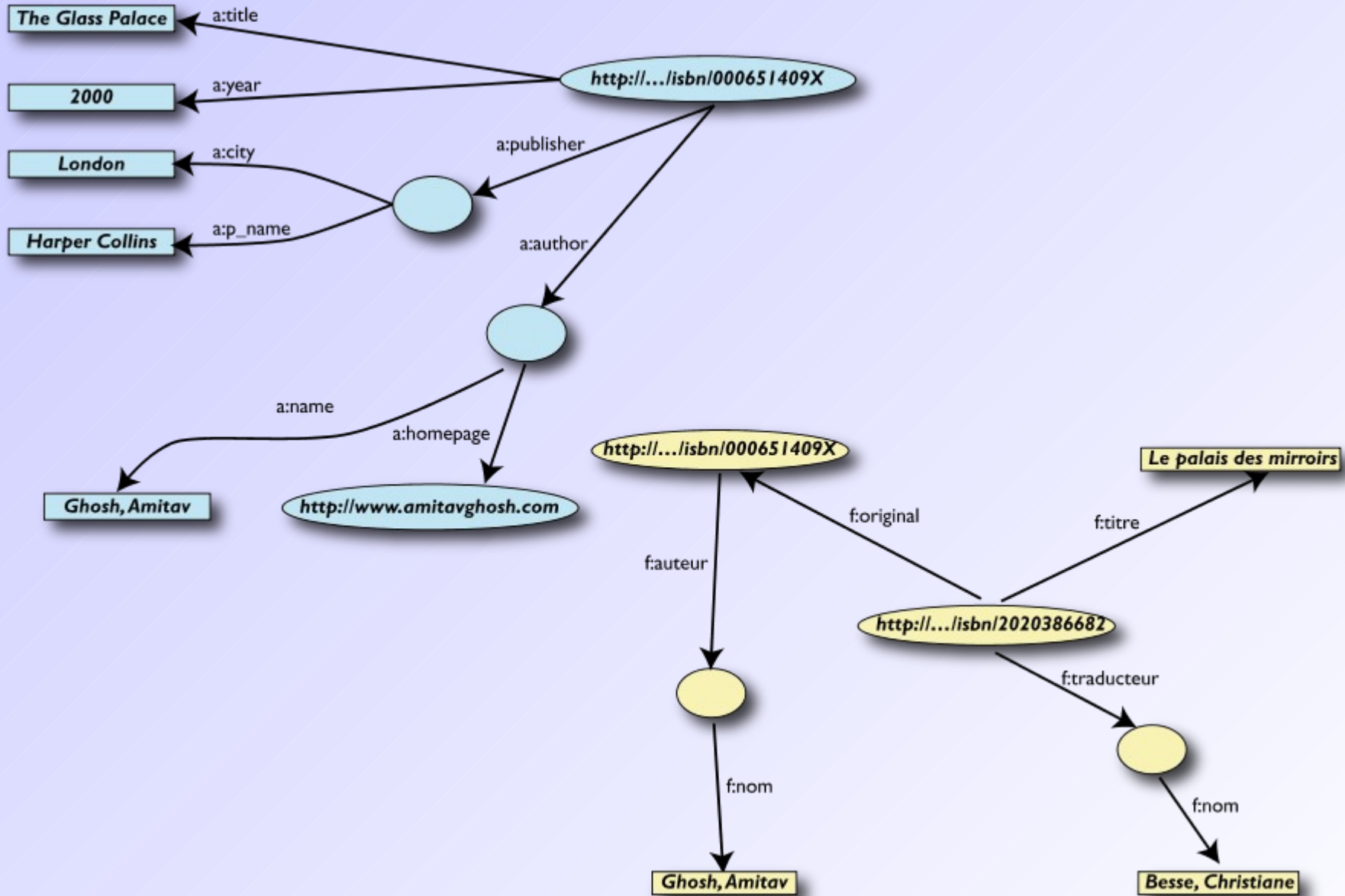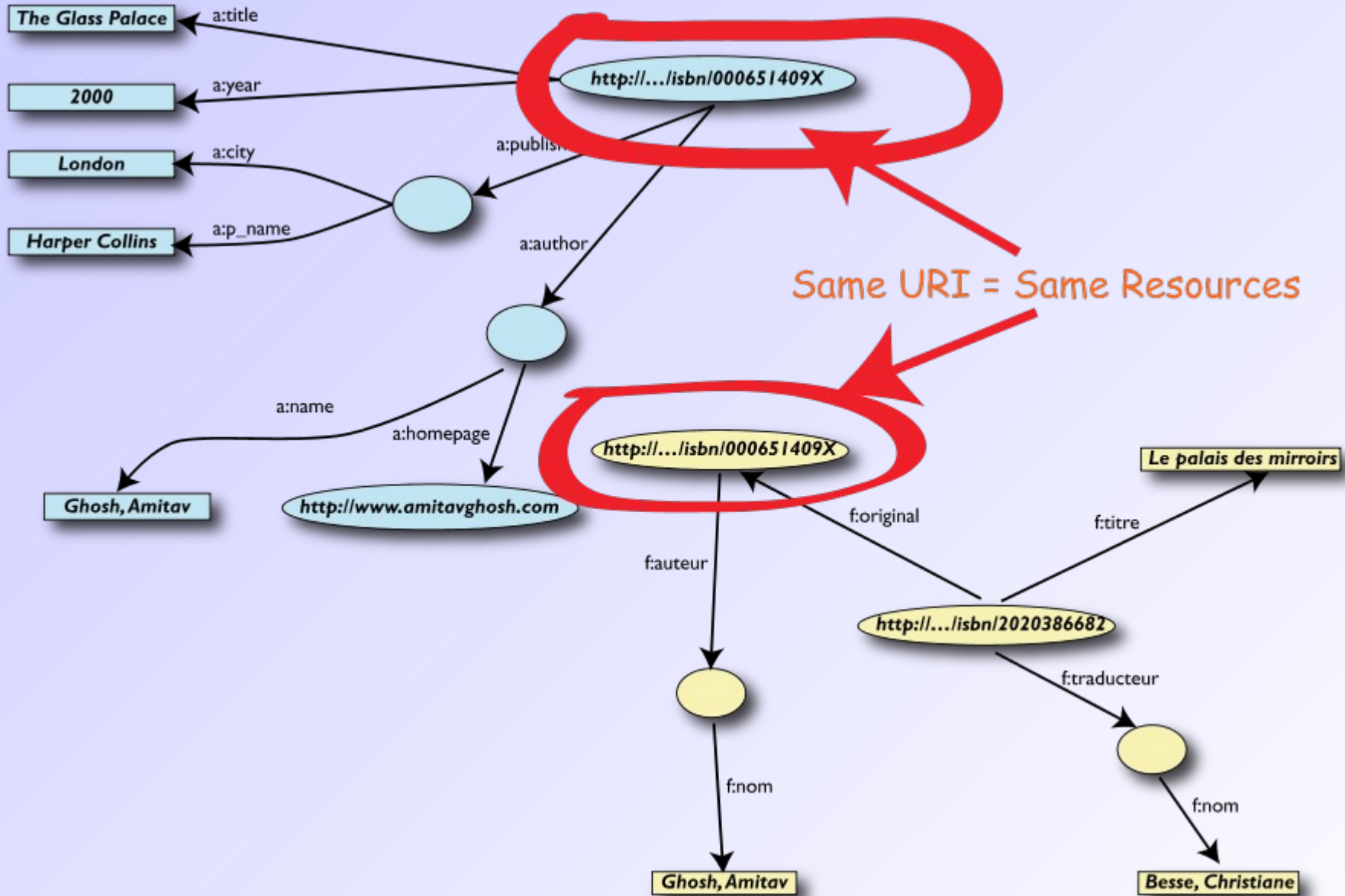
Merging identical nodes
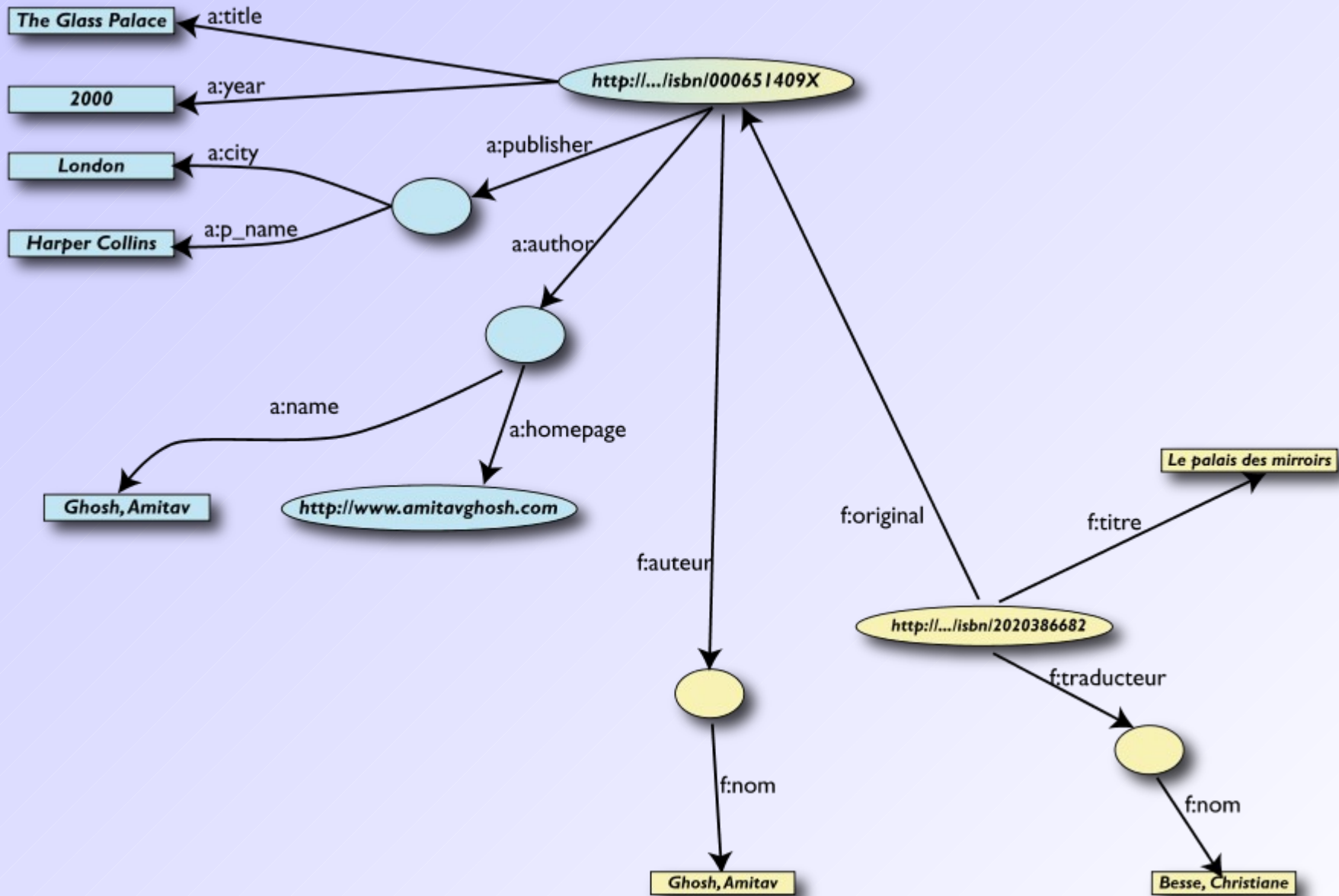
Add some missing knowledge

- We "feel" that a:author and f:auteur should be the same

- But an automatic merge doesn't know that without help

- We will 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
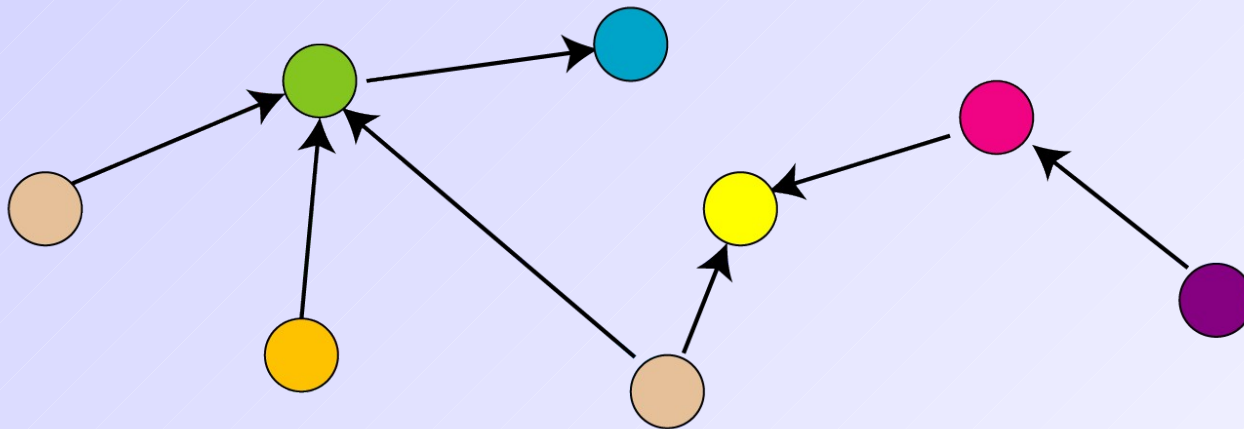 - it can be used as a "category" for certain type of resources

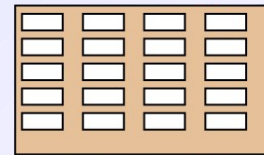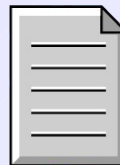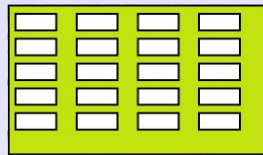# The merged relations
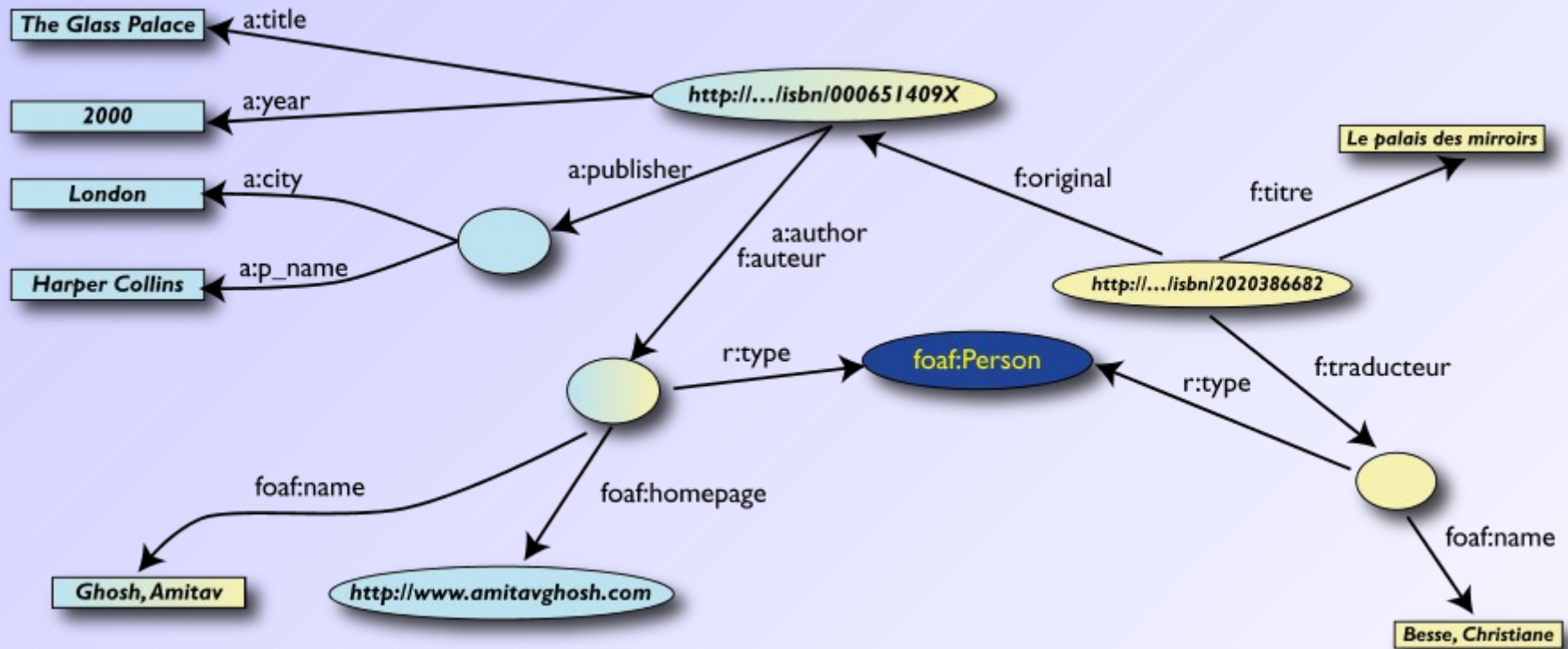
What did we do?
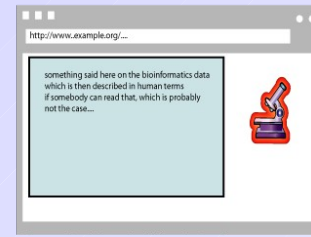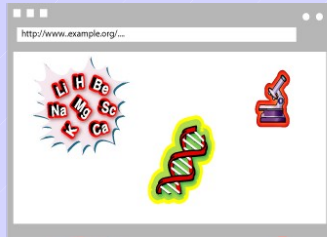
Applications

Query,
Manipulate,
etc.

Data represented in abstract format

Map,
Expose,
etc.

Data in various formats

Web of Data

- We should publish data on servers
  - In standard ways rather than ad hoc approaches
    - To encourage shared semantics for comparability and aggregation
- Set RDF links among the data items from different data sets
  - URIs as globally unique names
  - URIs for downloadable datasets (RDF graphs)
  - URIs for Web APIs including Sparql queries
- Encourage people to innovate
  - More data
  - More applications
- *Watch the network effect work its magic!*
  - *The value scales as the square of the number of participants*
    - *As described by Beckstrom, Reed, Metcalfe and originally Vail*

Linked Open Data Cloud, March 2008
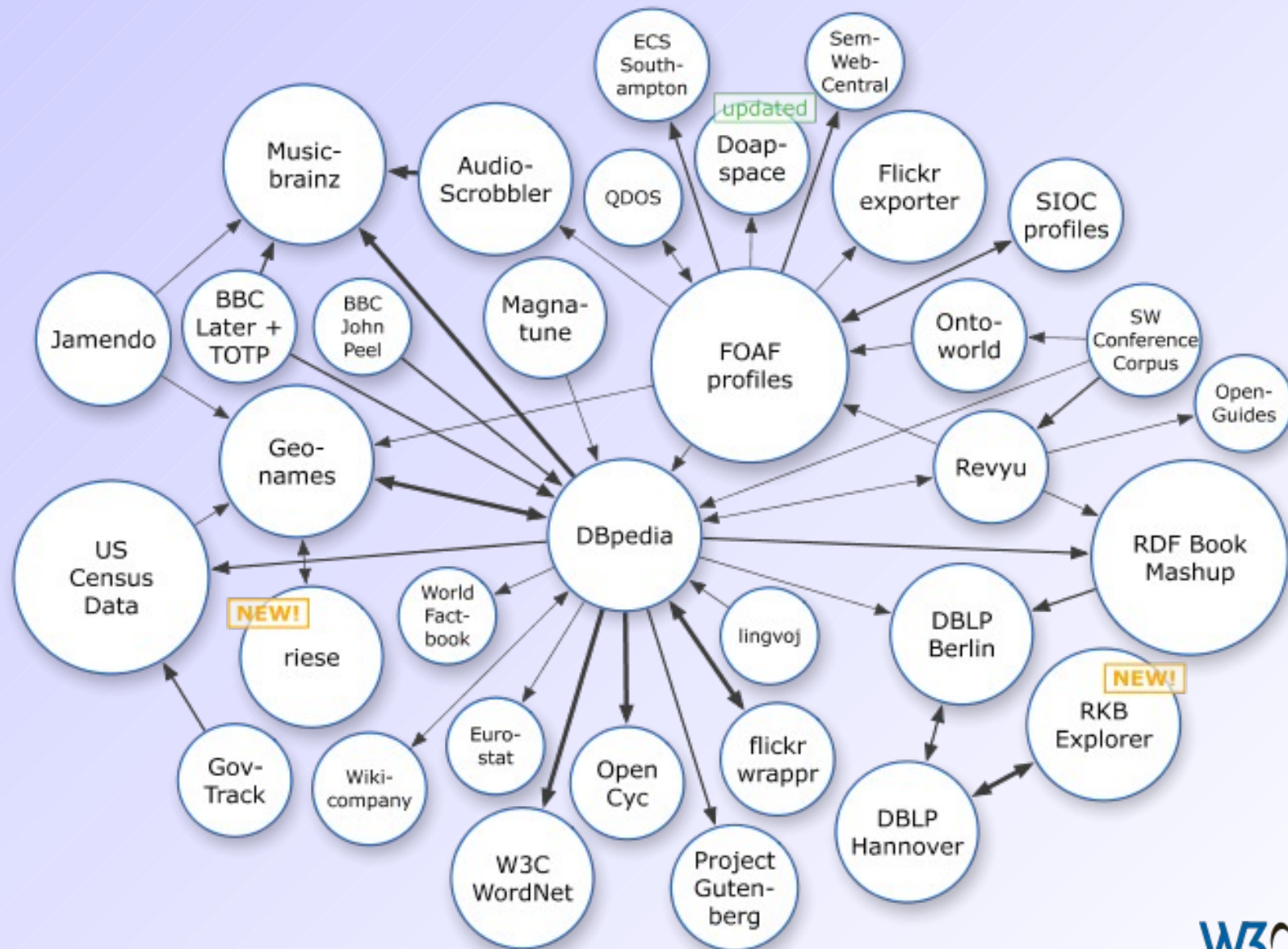
Linked Open Data Cloud, March 2009
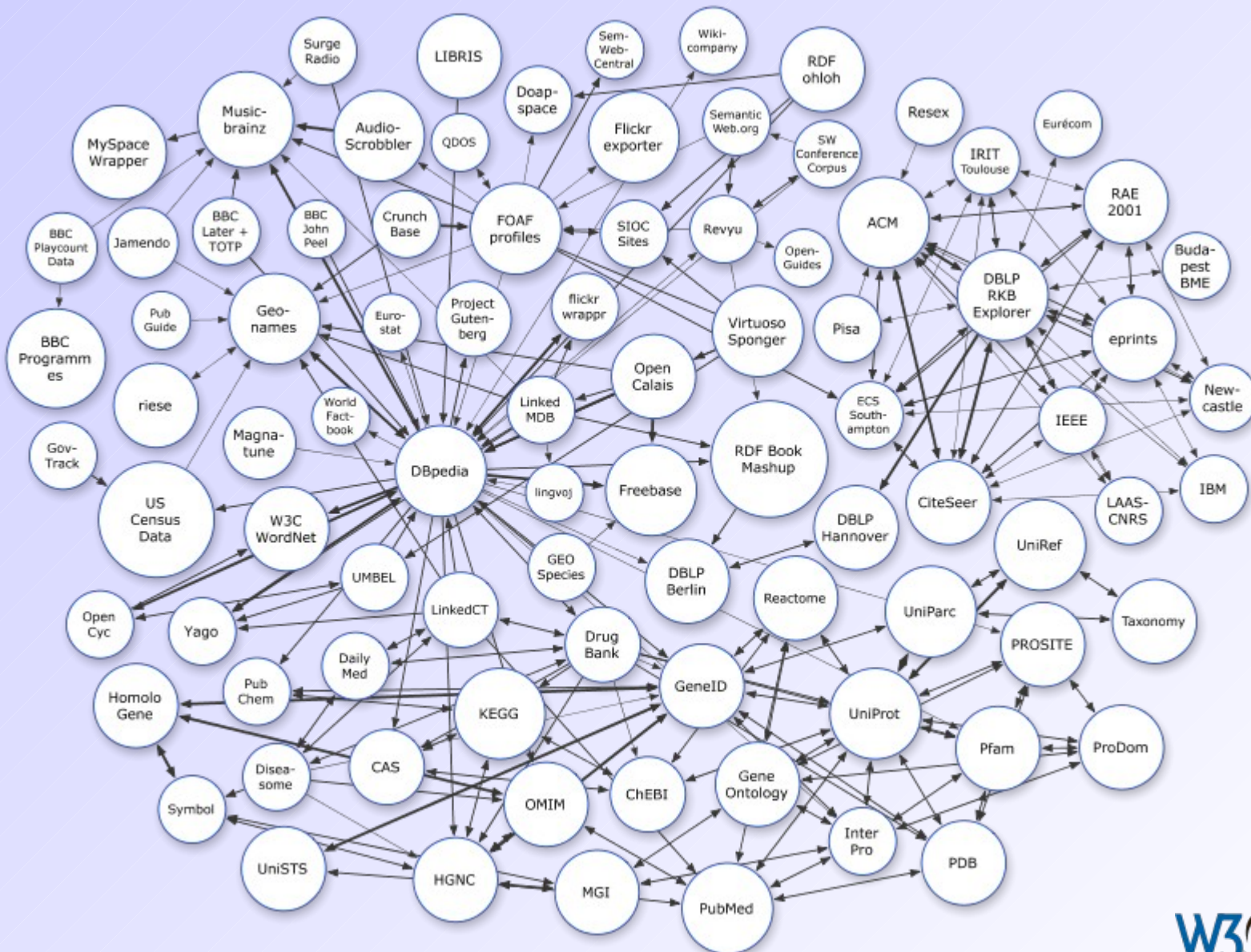
Corporate adoption

- Major companies offer (or will offer) Semantic Web tools or systems using Semantic Web: Adobe, Oracle, IBM, Software AG, GE, Northrop Gruman, Altova, Microsoft, Dow Jones, …

- Others are using it (or consider using it) as part of their own operations: Novartis, Pfizer, Telefónica, …

- Some of the names of active participants in W3C SW related groups: ILOG, HP, Agfa, SRI International, Fair Isaac Corp., Oracle, Boeing, IBM, Chevron, Siemens, Nokia, Pfizer, Sun, Eli Lilly, …

Query languages

Querying RDF with SPARQL

- A query language for RDF data

- Similar in syntax and spirit to SQL

```
SELECT ?p
  WHERE {
    ?L1 arcrole:parent-child ?b1 .
    ?b1 xl:type xl:link .
    ?b1 xl:from ?p
    OPTIONAL {
        ?L2 arcrole:parent-child ?b2 .
        ?b2 xl:type xl:link .
        ?b2 xl:to ?p
    }
    FILTER (!BOUND(?b2))
  }
```

Defining shared vocabularies

Data Types

- RDFS defines some predicates for common datatypes, e.g.
  - Booleans
  - Numbers
  - Strings
    - As XML or as natural language, e.g. Spanish
  - Dates
  - Classes
- Resources can belong to several classes

rdfs:Class

rdf:type

#Novel

rdf:type

http://.../isbn/000651409X
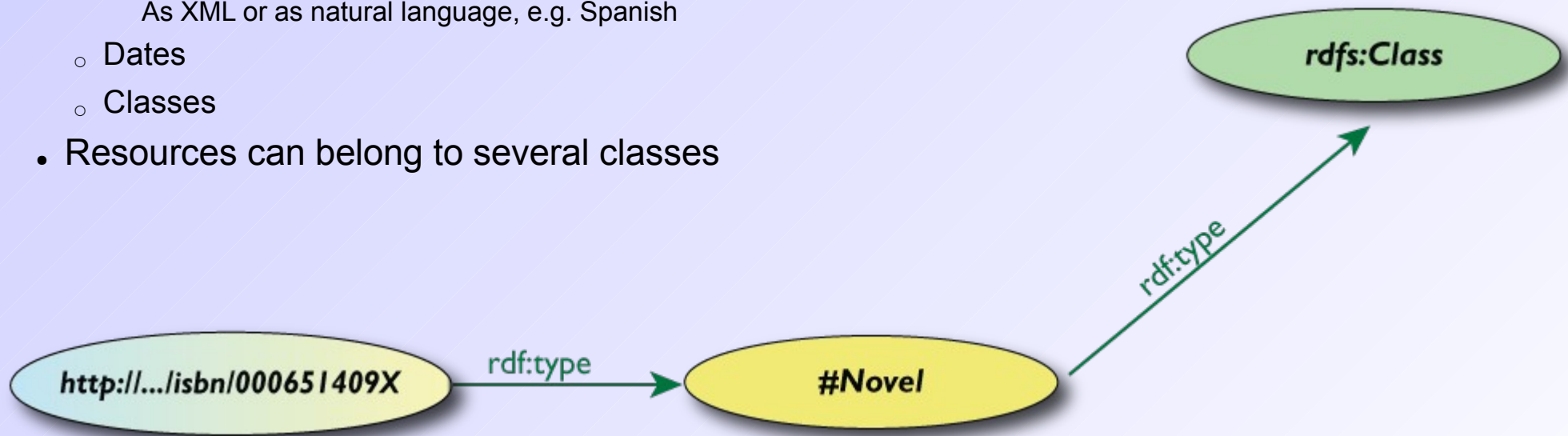
OWL for Ontologies

- RDFS is useful, but complex applications may want more

- OWL adds lots of possibilities
    - Characterization of properties
    - Disjointness or equivalence of classes
    - In RDFS, you can subclass existing classes
    - In OWL, you can construct classes from existing ones
        - Through set intersection, union, complement, etc.

- But this comes at a cost...

OWL Profiles

- Trade off between rich semantics for expressibility and ease of making inferences

  - Simpler inference engines are possible with restrictions on which terms can be used and under what circumstances

- OWL full

  - Very expressive, but not computable in general

- OWL DL

  - Popular computable subset of OWL full

- OWL 2 defines further profiles

Business Rules

Rule Languages

- May be more convenient than ontologies

- Example

  - A cheap book is a novel with over 500 pages and costing less than $8

- W3C Rule Interchange Format (RIF)

  - Family of languages for rule interchange

    - For different kinds of rule language

  - Uses include

    - Negotiating eBusiness contracts across platforms

    - Access to business rules of supply chain partners

    - Managing inter-organizational business policies

W3C Semantic Web

XBRL and the Semantic Web

XBRL – an XML format for company reports where each reported fact is tagged with its context in the reporting taxonomy, e.g. US GAAP or IFRS

XBRL
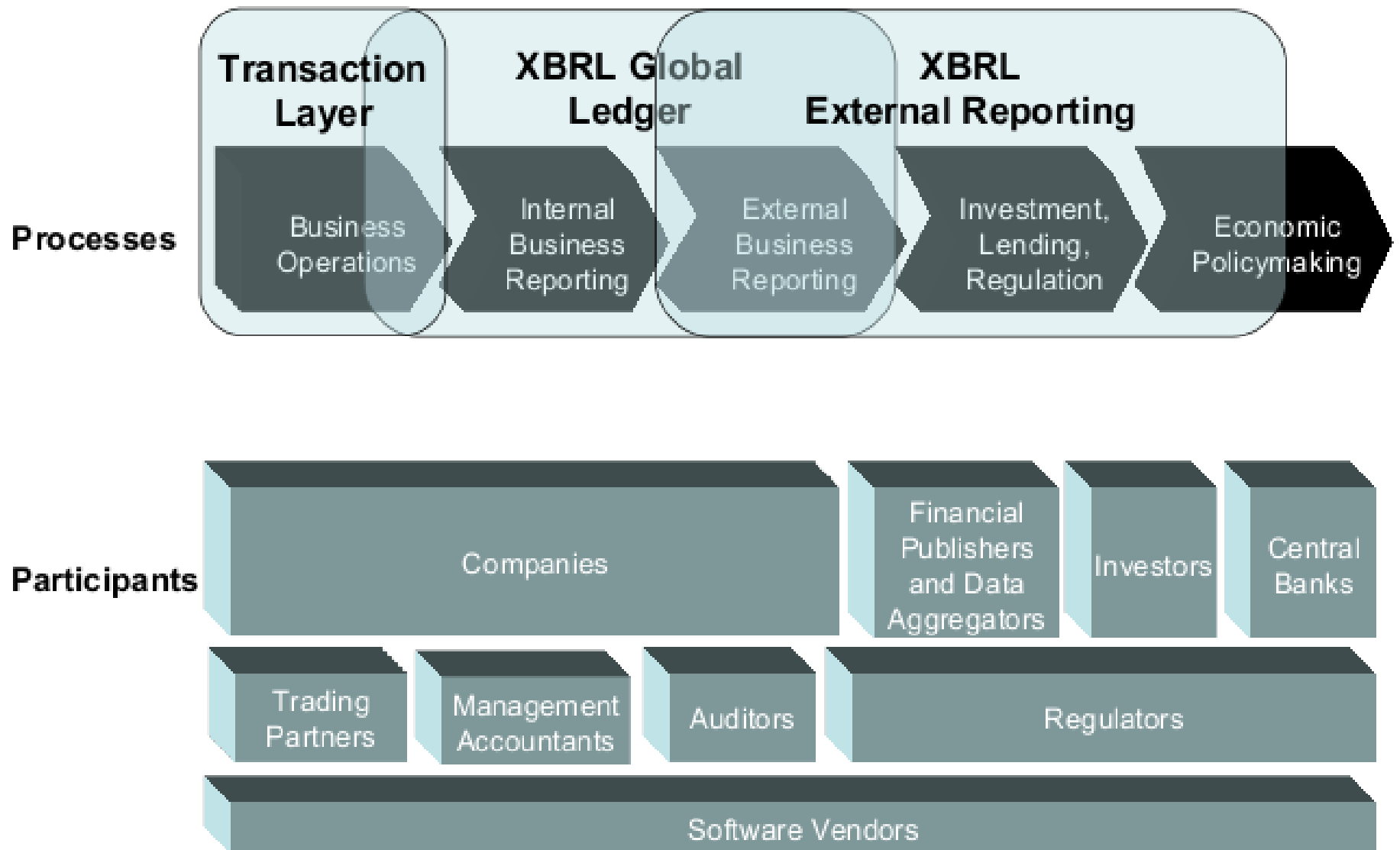
Why translate XBRL to another format?

- It is very expensive to process 10-50MB of XML on each query

  - Memory and CPU intensive: about one second of CPU time per 10MB of XML source

- Better to pre-process filings into a persistent format designed to match needs of queries

  - Current tools use proprietary solutions

- RDF and OWL as natural choices

  - Mature standards

  - Facilitate mashing financial data with other kinds of information available over the Web

  - Web APIs and standards would enable an ecosystem of value adding players

XBRL as RDF/Turtle

Part of US GAAP taxonomy

```
@prefix usfr-pte: <http://www.xbrl.org/us/fr/common/pte/2005-02-28>.

usfr-pte:ChangeOtherCurrentAssets
    rdf:type xbrli:monetaryItemType;
    xbrli:periodType "duration".
usfr-pte:ChangeOtherCurrentLiabilities
    rdf:type xbrli:monetaryItemType;
    xbrli:periodType "duration".

_:link155 arcrole:parent-child [
    xl:type xl:link;
    xl:role role1:StatementFinancialPosition;
    xl:use "prohibited";
    xl:priority "1"^^xsd:integer;
    xl:order "1.0"^^xsd:decimal;
    xl:from usfr-pte:IntangibleAssetsNetAbstract;
    xl:to usfr-pte:IntangibleAssetsGoodwill;
    ].
```

W3C SemanticWeb

XBRL as RDF/Turtle

Sample of an XBRL Instance file

```
_:context_FY07Q3
    xl:type xbrli:context;
    xbrli:entity [
        xbrli:identifier "0000789019";
        xbrli:scheme <http://sec.gov/CIK>;
        ];
    xbrli:period (
        [ xbrli:startDate "2007-01-01"^^xsd:date;
          xbrli:endDate "2007-03-31"^^xsd:date; ]
        ).

_:unit_usd xbrli:measure iso4217:USD.

_:fact209
    xl:type xbrli:fact;
    xl:provenance _:provenance1;
    rdf:type us-gaap:PaymentsToAcquireProductiveAssets;
    rdf:value "461000000"^^xsd:integer;
    xbrli:decimals "-6"^^xsd:integer;
    xbrli:unit _:unit_USD;
    xbrli:context _:context_FY07Q3.
```

XBRL and OWL

- XBRL Taxonomy loosely equates to OWL ontology
  - But note XBRL's taxonomy overrides
- Automated mapping is mostly feasible
  - As demonstrated by Rhizomik XSD2OWL
- XBRL's formal semantics are weak
- XBRL versioning standard will describe differences between different versions of the same taxonomy, e.g. US GAAP 2008, 2009
  - Unaware of work on mapping this into OWL
  - Is it a good match to real world needs?
    - e.g. rules of thumb for computing analytic ratios
- Reasoning across different taxonomies remains a major challenge
  - e.g. US GAAP vs IFRS
- Need for standards for business & financial data that are syntax independent

Web-based ecosystem
for financial information

- Publishers of raw data
  - Investor relation websites
  - Government agencies
  - News agencies
- Data aggregators
  - Republish data as linkable triples, Sparql queries
  - Higher level APIs for common queries
    - Results as charts or tables
  - Web of scripts that add value
    - Custom analytics across filings
    - Export triiples, high level APIs or presentations
- Smart search engines
- Communities
  - Share reviews, comments, analyses, mashups, ...

Smart Search Engines

- Imagine search engines that provide selected financial highlights for each company that matches the search criteria you just entered

  - With salient numbers and charts

- The search results tailor the data provided according to your interests

  - Based upon analysis of the search criteria and other information gleaned from previous searches

  - Subject to your privacy preferences, of course! **

- Interactive data you can drill down on

- Search engines can also be used within Intranets!

** My other job is on privacy and identity management for an EU FP7 project

Summary

- The Web succeeds by connecting people
  - The power of the network effect!
- The Web of data
  - Rich models of concepts and relationships
  - Access to data and meta-data as basis for comparability
  - Rules of thumb for overcoming variations
  - Hiding internal representations and APIs
- Web-based ecosystem for financial information
  - Many ways to add value building on the work of others
  - Semantic Web as solution to data integration
- What's needed to drive this forward?

Thank you for listening

Dave Raggett, W3C

dsr@w3.org