LTI
Language, Technology and the Internet

The Semantic Web

Francis Bond
Division of Linguistics and Multilingual Studies
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Lecture 10

Revision of Text and Meta-text

Text and Meta-text

- Explicit Meta-data
  - Keywords and Categories
  - Rankings
  - Structural Markup
- Implicit Meta-data
  - Links and Citations
  - Tags
  - Tables
  - File Names
  - Translations

Explicit Metadata

- You can get information from metadata within documents
  - When they are accurate they are very good
  - They are often deceitful
- HTML, PDF, Word, ...Metadata
- Keywords and Tags
- Rankings
- Links and Citations
- Structural Markup

Implicit Metadata

- You can get clues from metadata within documents
  - as they are non-intended, they tend to be noisy
  - but they are rarely deceitful
- HTML tags as constituent boundaries
- Tables as Semantic Relations
- File Names (content type and language)
- Translations — Bracketed Glosses; Cross-lingual Disambiguation
- Query Data
- Wikipedia Redirections and Cross-wiki Links

The Semantic Web

The Semantic Web

- What is it?
- How is it built?
- Why is it being built?
- Problems

The Semantic Web

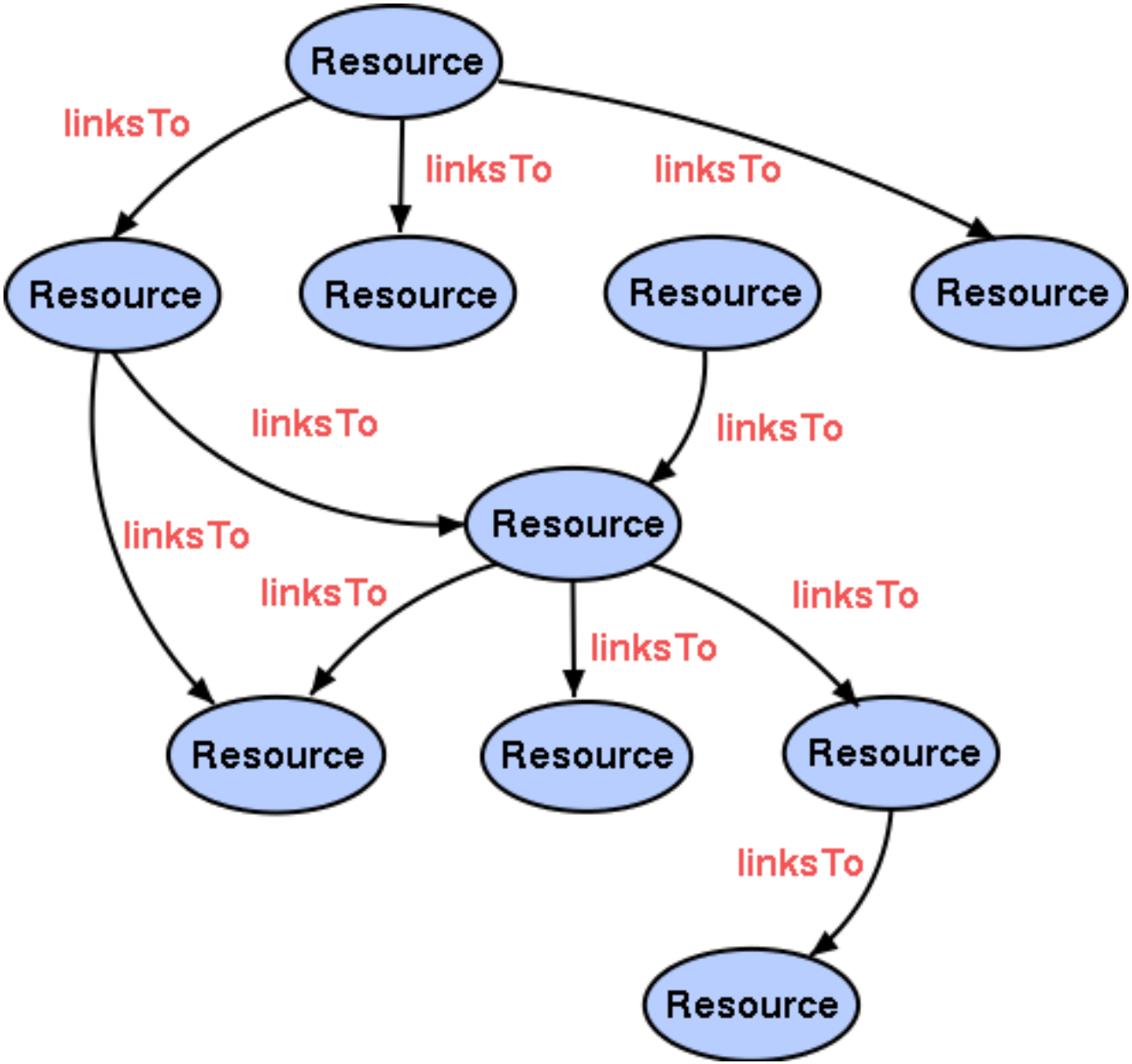
- A vision of a more useful World Wide Web
- the meaning of information and services on the web is defined
- machines (and people) can understand the web

The Web as a universal medium for
data, information, and knowledge exchange.

The Web as it is now

- Resources
  - Identified by URLs (uniform resource locator)
  - Untyped
- Links
  - href, src, ...
  - limited, non-descriptive
- Human User
  - Exciting World - Semantics deduced from context
- Machine User
  - Very little explicit information

The Current Web



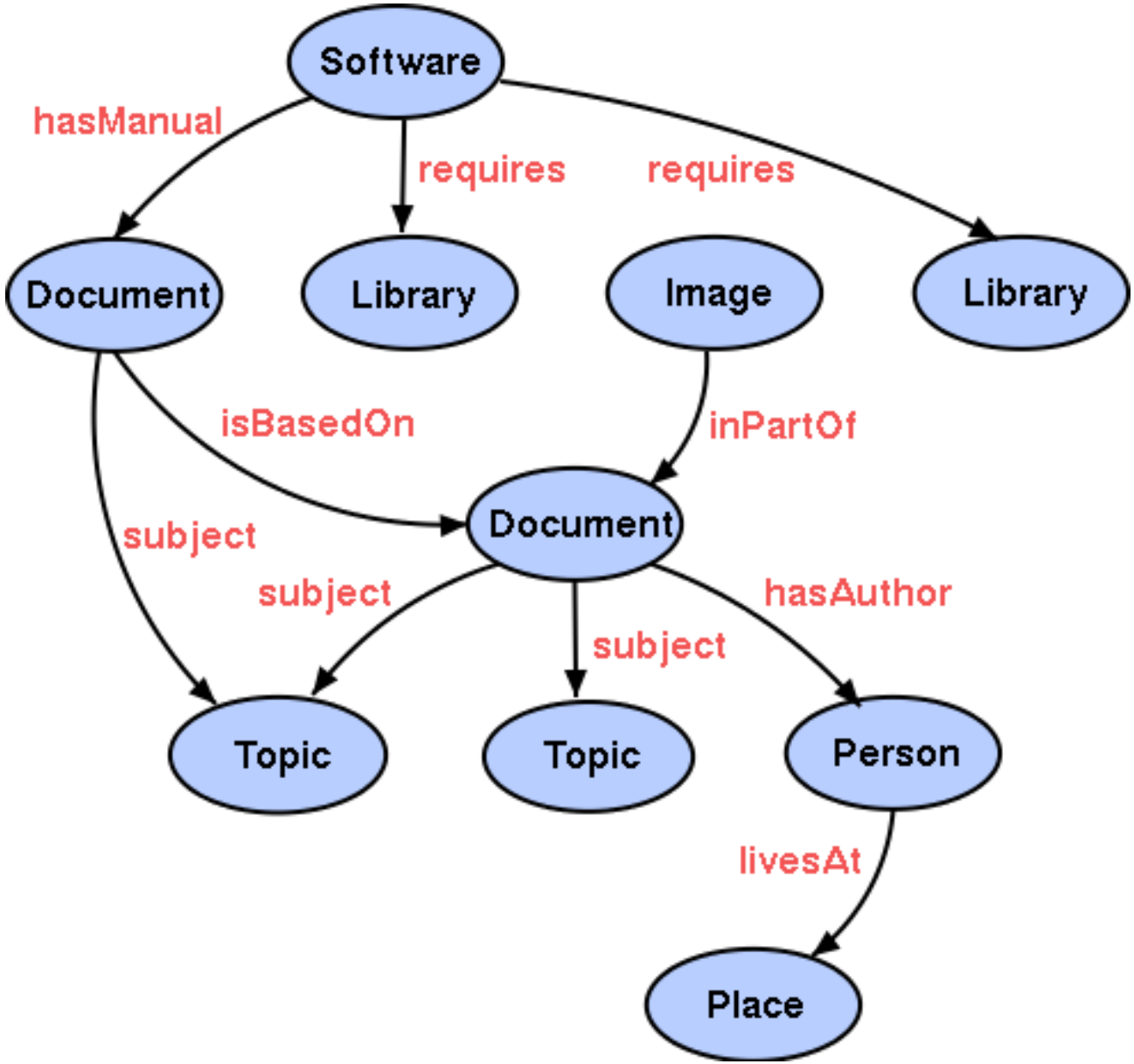
The Web as it could be

- Resources
  - Globally Identified by URIs (uniform resource identifier)
  - Extensible, Relational
- Links
  - Identified by URIs
  - Extensible, Relational

Use the web to define the web

- Human User
  - Even more exciting world - Richer user experience
- Machine
  - More processable information

The Semantic Web

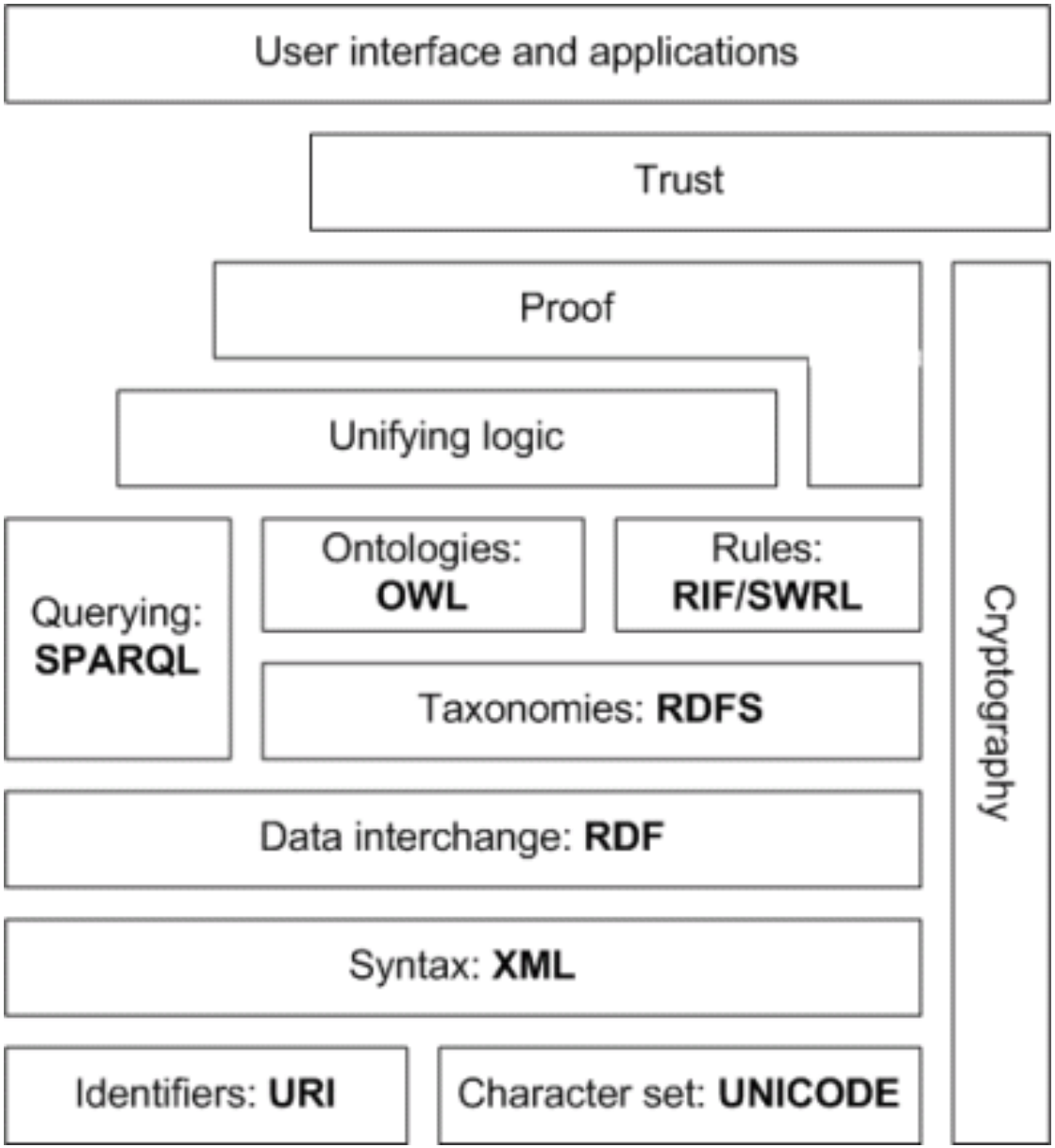


Semantic Web Goals

- Web of Data
  - provides common data representation framework
  - makes possible integrating multiple sources
  - so you can draw new conclusions
- Increase the utility of information by connecting it to definitions and context
- More efficient information access and analysis

E.G. not just "color" but a concept denoted by a Web identifier:
<http://en.wikipedia.org/wiki/Sapphire_(color)>

Semantic Web Architecture



Identifiers and Characters

- Characters are always defined using Unicode
- Identifiers are Uniform Resource Identifiers
  - Universal Resource Name
    - e.g., `urn:isbn:1575864606`
    - uniquely identifies one edition of a book
  - Universal Resource Locator:
    - `scheme:part?query#anchor`
    - e.g., `http://www3.ntu.edu.sg/home/fcbond/`
    - http, skype, ssh, secondlife, ...
- Identifies or names a resource on the web

XML: eXtensible Markup Language

- XML is a set of rules for encoding documents electronically.
- Based on a simplified SGML
- XML' s design goals emphasize simplicity, generality, and usability.
- It is a textual data format
- It supports many encodings, with Unicode preferred
- It can represent arbitrary data structures, for example in web services.

In other words

- XML stands for EXtensible Markup Language
- is a markup language much like HTML
- was designed to carry data, not to display data
- tags are not predefined. You must define your own tags
- is designed to be self-descriptive
- XML is a W3C Recommendation

With XML You Invent Your Own Tags

```
<bitext>
<author xml:lang="eng">Francis Bond</author>
<author xml:lang="jpn">フランシス ボンド</author>
<email>fcbond@ntu.edu.sg</email>
<opendata>yes</opendata>
```

- The tags describe the content, but do not do anything.
- You have to define the meaning elsewhere

XML structure

- An XML element looks like this

```
<tag attribute='value'>Content</tag>
```

  - You can use attributes (easy to verify)

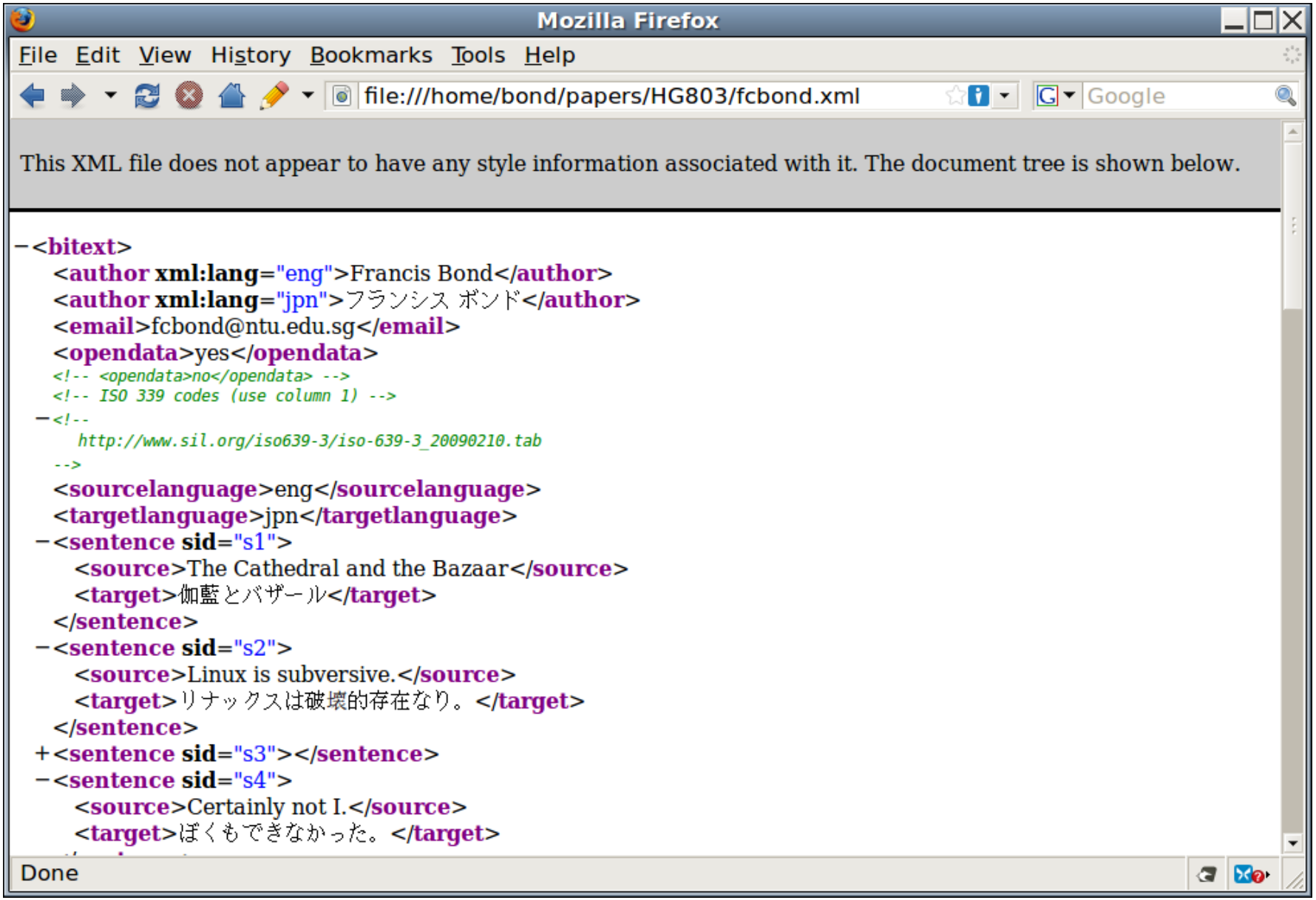
```
<author xml:lang="eng">Francis Bond</author>
```

  - or nested elements (flexible)

```
<author>
  <lang>eng</lang>
  <name>Francis Bond</name>
</author>
```

- If all elements are written correctly the document is well formed

XML with structure highlighted



Document Type Definitions

- You can also define what tags are possible with
  - DTD (Document Type Definition)
  - XML Schema

```
<!DOCTYPE bitext
[
<!ELEMENT author (#PCDATA)>
<!ATTLIST author xml:lang CDATA>
<!ELEMENT sentence (source,target)>
...
```

- This makes the structure explicit

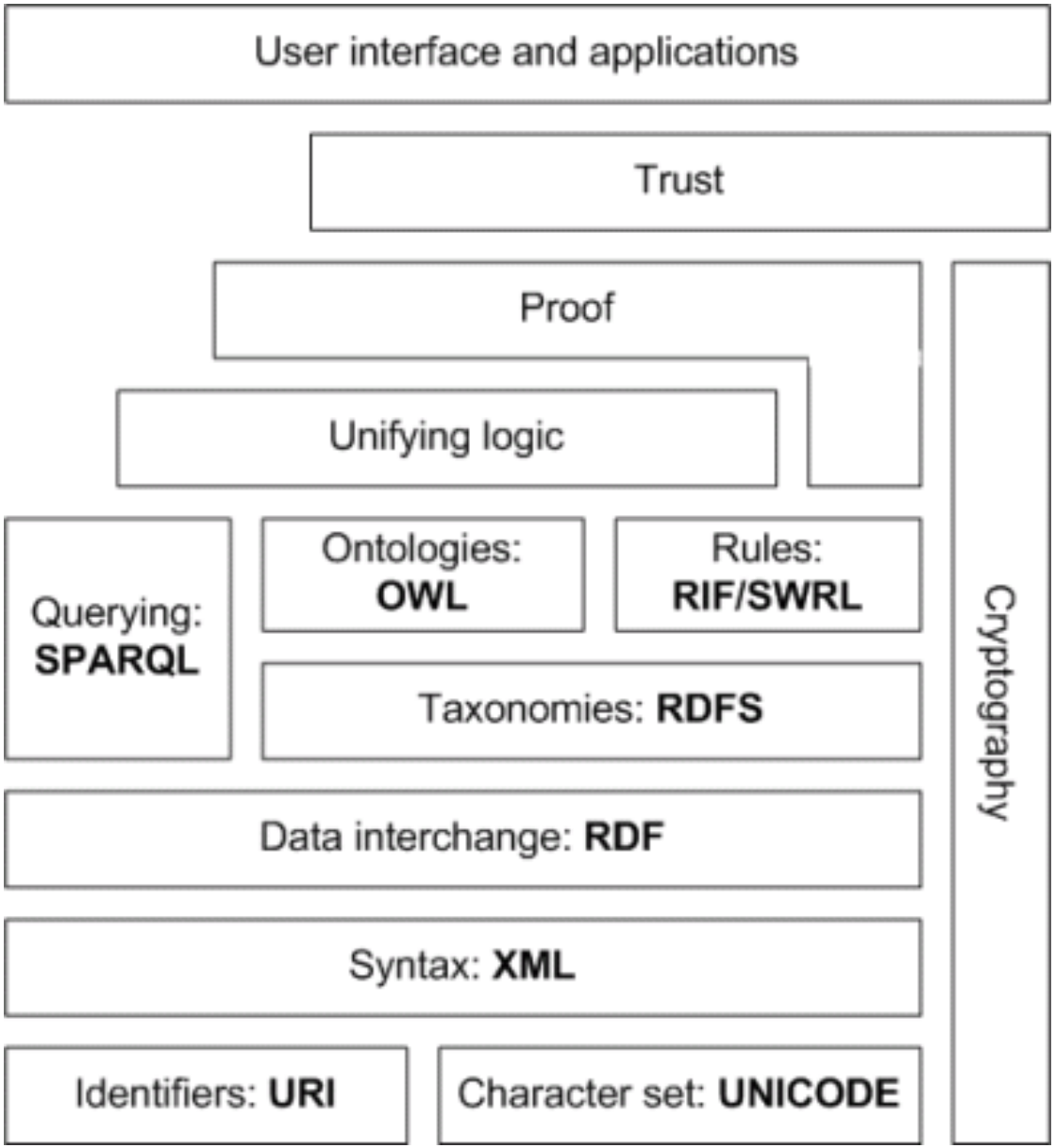
Why Use a DTD?

- With a DTD, XML files can carry a description of their own format.
- With a DTD, independent groups of people can use a well defined, enforceable standard for interchanging data.
- Your application can use a standard DTD to verify that the data you receive from the outside world is valid.
- You can also use a DTD to verify your own data.
  - You can catch format errors early
- **BUT**, XML only defines structure not content

Validation

- Validation is very important
- Ill-formed data makes parsing complex
- Early detection of errors is cost-effective
- Validated data is easy to maintain

Semantic Web Architecture



RDF: Resource Description Framework

- We want to identify content
- Annotate information with descriptions:
  triples of information
  - subject
  - `predicate`
  - object

  e.g. <dog, `hyponym`, animal>
  e.g. <Francis Bond, `teacher`, HG252>
- The trick is that each element is a URI
- You can say anything about anything

RDF graph describing Eric Miller

Relational Semantic Representation (again!),
with relations defined by URIs

RDFs are written using XML

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">

  <contact:Person rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:fullName>Eric Miller</contact:fullName>
    <contact:mailbox rdf:resource="mailto:em@w3.org"/>
    <contact:personalTitle>Dr.</contact:personalTitle>
  </contact:Person>

</rdf:RDF>
```

- Not designed for people to read (or write) directly
- Designed to be extensible and explicit

Why use URIs in RDFs with XML?

- A shared URI gives a shared definition
- You can add new URIs as necessary
- RDFs are complicated
  - you need to validate the syntax $\rightarrow$ XML
- RDFs assume the existence of the web
  - Nothing has meaning on its own
  - *You shall know a URI by the company it keeps*

OWL and Ontologies

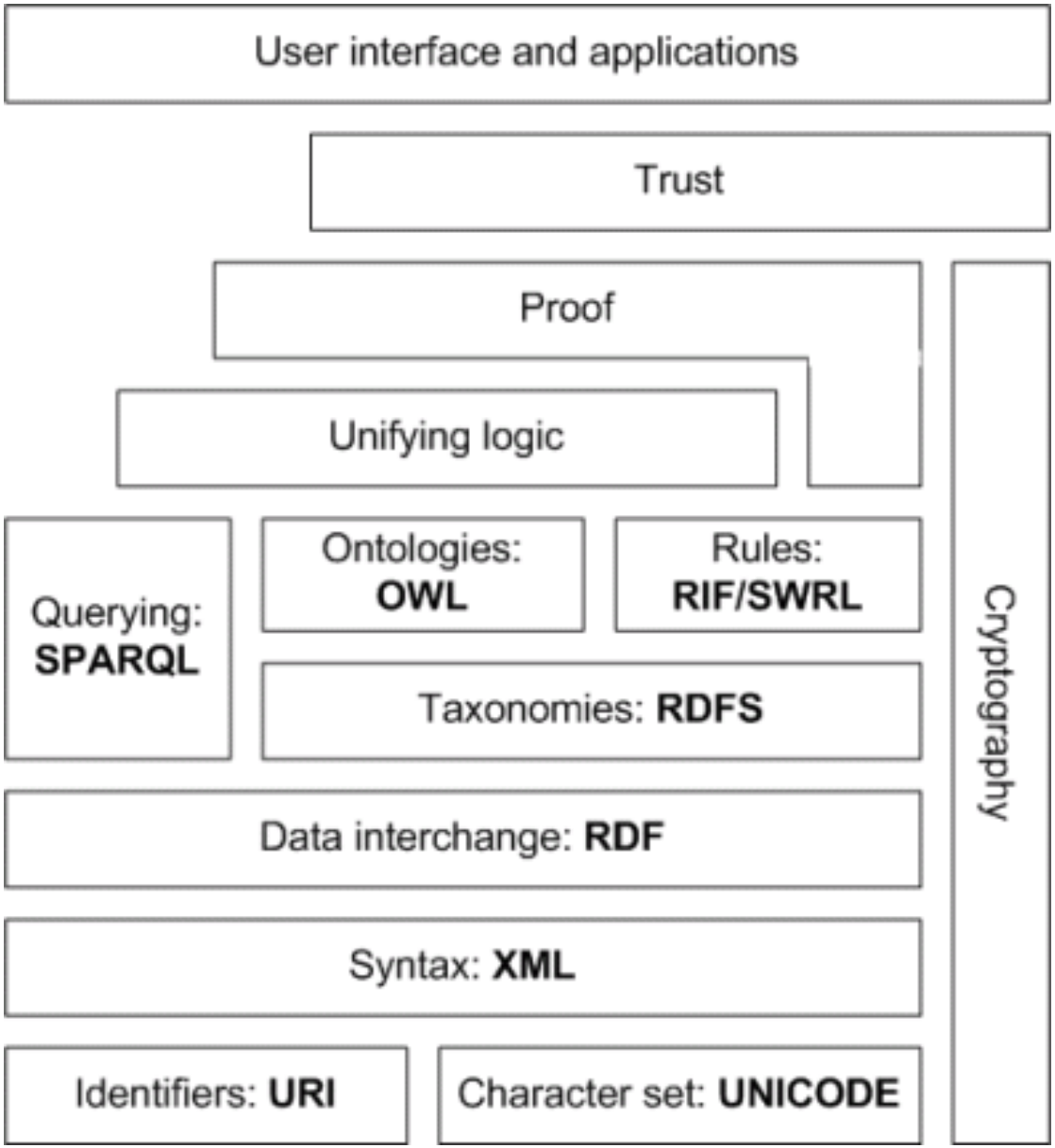
- Allowing any URI makes information hard to combine
- Use Ontologies to link it together again
  - You can normally agree on a hypernym (supertype)
- Agreeing on an ontology is difficult
  - Many detailed ontologies
  - One common ontology
    - The Standard Upper Merged Ontology (SUMO)
    - Links many ontologies (including WordNet)
  - Much recent work on medical and library domains

Common Semantic Web Ontologies

- People + Organisations
  - FOAF, HCard, Relationship, Resume
- Places
  - Geonames, Geo
- Events
  - RDFCalendar
- Social Media
  - SIOC, Review
- Topics + Tags
  - SKOS, MOAT, HolyGoat

- eCommerce
  - GoodRelations, CC Licensing
- More...
  - Scovo, DOAP, Recipes, Measurements, ...
- General things
  - SUMO, WordNet

Semantic Web Architecture



The upper layers

- Querying the (semantic) web is
  - Slow (non local access $\rightarrow$ millions of times slower)
  - Non-deterministic (the answers change)
  - Vast (could be trillions of elements)

  A whole new set of technical problems

- Logic, Proof and Trust are still works in progress
  - Reusing AI research from the 70s and 80s

- But applications are going along without them
  - Enhanced search using metadata
  - Combinations of data

Semantic Web Lessons

- RDF as a general information model is applicable to many uses
  (many of which we never even thought about)
- Common data representation and architecture drives down costs
  (technical and social)
- Facilitates serendipitous interoperability
  - breaking down the barriers of domain knowledge
- When "Anyone can say anything about anything",
  who you trust is important (same as in text mining)
- Beneficial to solving interoperability in Open (rather than Closed) systems

Summary

- The goal of the Semantic Web is to share knowledge
  - Uses markup to give tractable annotation
    - Unicode
    - XML
    - URI
    - RDF
    - OWL
  - Relies on web resources to make common assumptions explicit

Example of making information explicit

< Francis Bond, author, Translating the Untranslatable >

- Francis Bond: http://www3.ntu.edu.sg/home/fcbond/
  - defined using Francis Bond's homepage
- author: http://purl.org/dc/elements/1.1/creator
  - defined using the Dublin Core Ontology

or http://nlpwww.nict.go.jp/wn-synset.cgi?synset=10794014-n

  - defined using WordNet

- Translating the Untranslatable: urn:isbn:1575864606
  - defined using the ISBN ontology

An Example of Data Integration

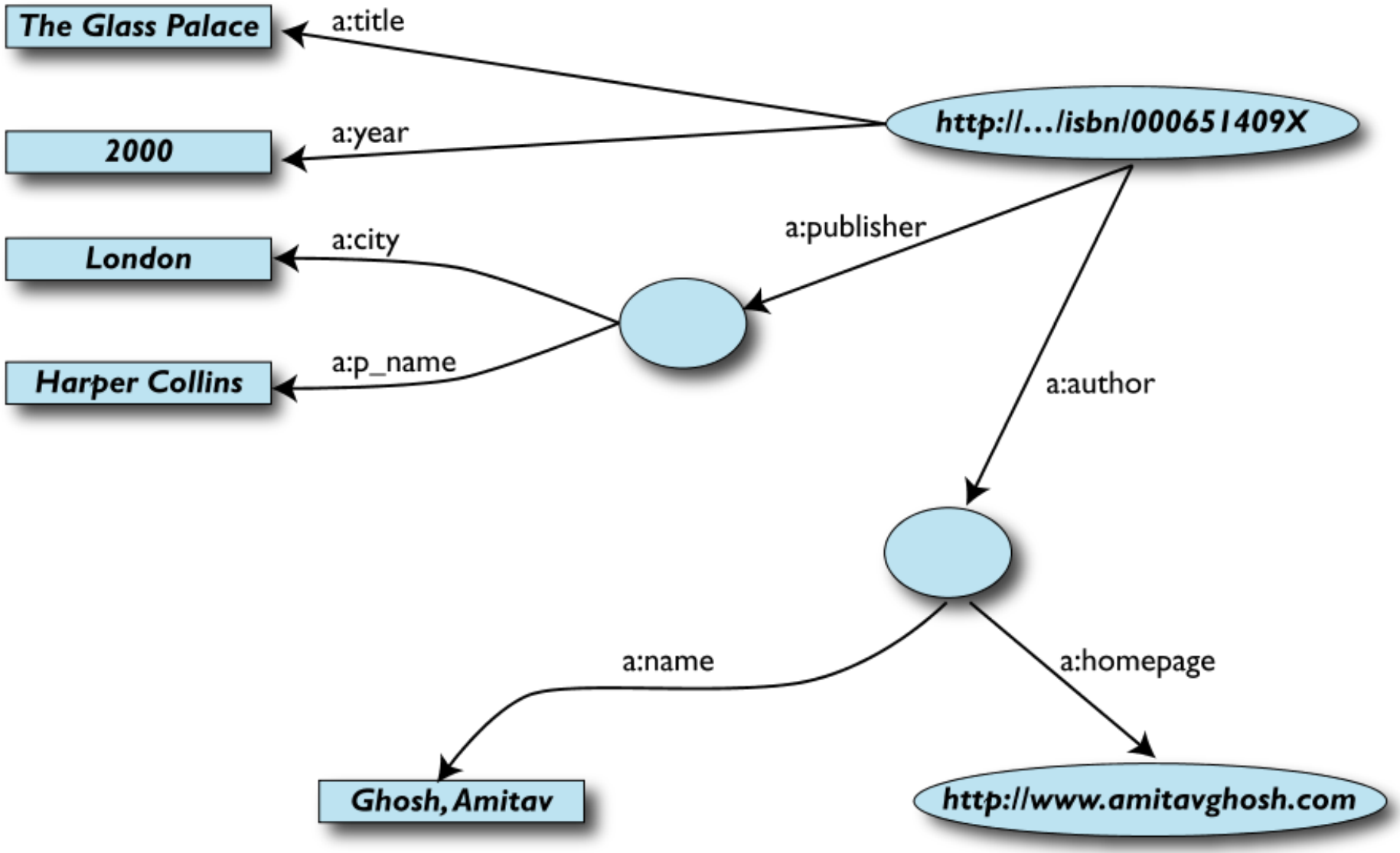
- Map the various data onto an abstract data representation
  - make the data independent of its internal representation
- Merge the resulting representations
- Start making queries on the whole!
  - queries that could not have been done on the individual data sets

Slides from "An introduction to the Semantic Web (Through an Example)"
By Ivan Herman (http://www.w3.org/People/Ivan/CorePresentations/IntroThroughExample/)

A simplified bookstore data (dataset "A")

- Book
  - ID: 000651409X
  - Author: A11
  - Title: The Glass Palace
  - Publisher: P202
  - Year: 2000
- Person:
  - ID: A11
  - Name: Amitav Ghosh
  - Homepage: www.amitavghosh.com
- Publisher:
  - ID: P202
  - Name: Harper Collins
  - Address: London

Export your data as a set of relations



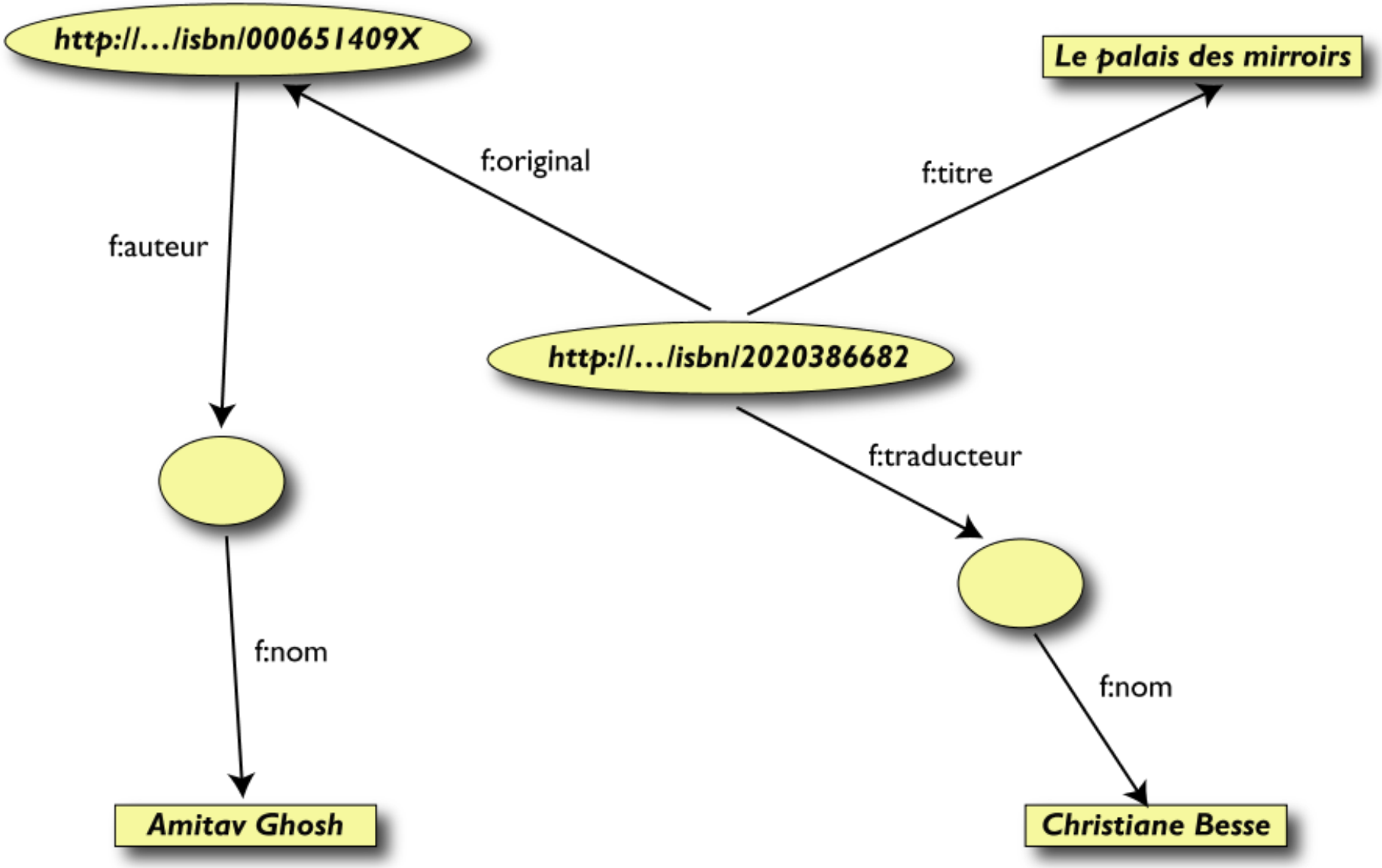
Some notes on the exporting the data

- Relations form a graph
  - the nodes refer to the "real" data or contain a string
  - 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
      - through text mining
      - etc.
    - One can export part of the data

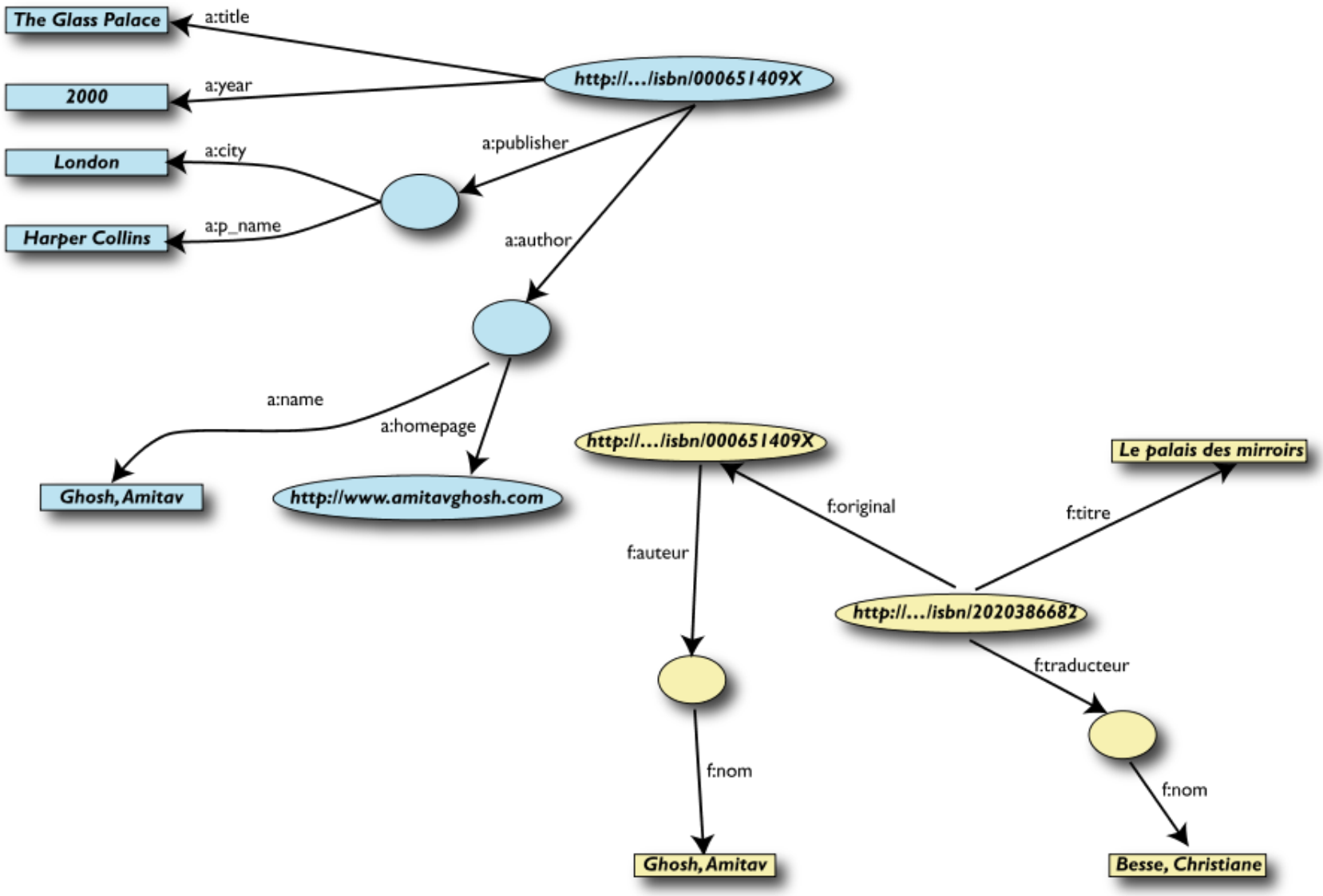
Another Set of Data (set "F")

- Book
  - ID: ISBN 20203866682
  - Titre: la Palais des miroirs
  - Traducteur: P2
  - Original: ISBN 000651409X
- Book
  - ID: 000651409X
  - Auteur: P1
- Person:
  - ID: P1
  - Nom: Ghosh, Amitav
- Person:
  - ID: P2
  - Nom: Besse, Christianne

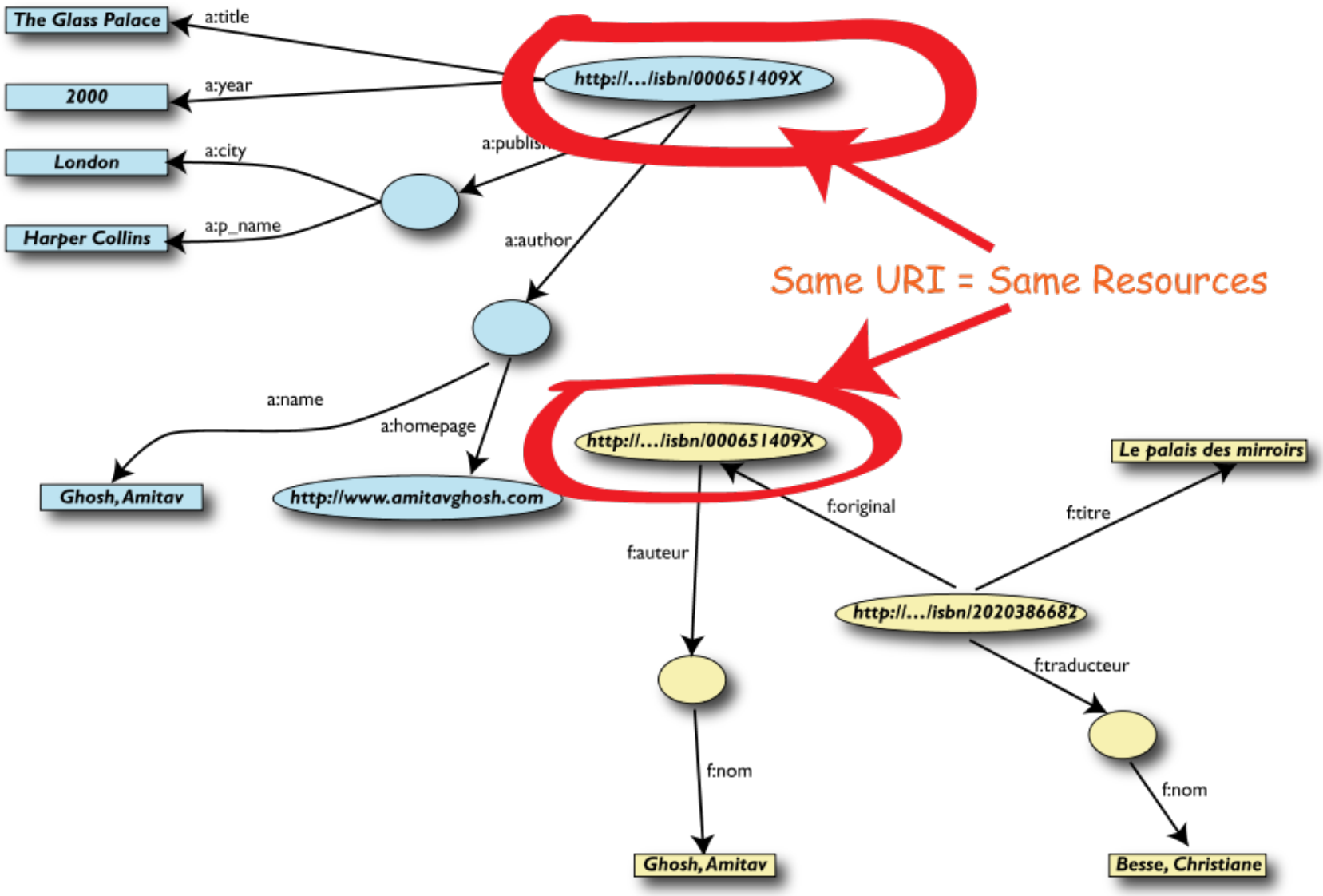
Export Set F


http://…/isbn/000651409X
Le palais des mirroirs
f:original
f:titre
f:auteur
http://…/isbn/2020386682
f:traducteur
f:nom
f:nom
Amitav Ghosh
Christiane Besse

Start to Merge



Merge Identical Resources



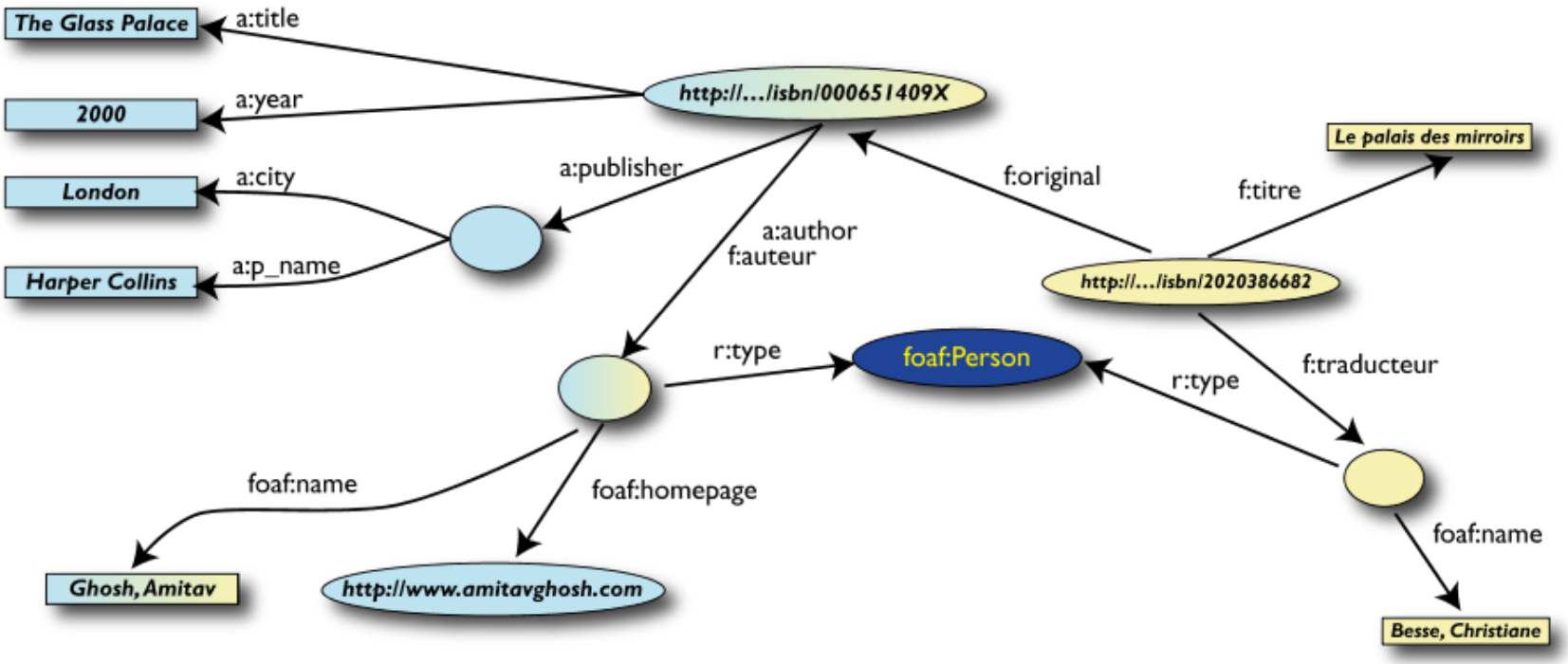
Start making queries

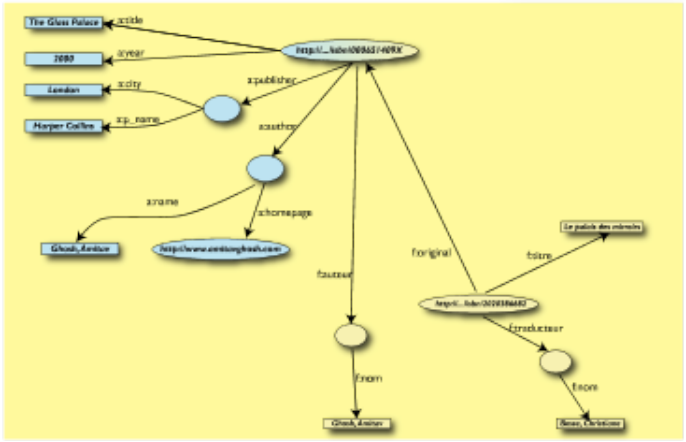
- User of data “F” can now ask queries like:
  - “give me the title of the original”
  - well, “donnes-moi le titre de l’ original”
- This information is not in the dataset “F”
- ...but can be retrieved by merging with dataset “A”!

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!
  - If only we could understand language
  - They both point to the same synset in wordnet (or at least one)
- 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

Use more knowledge

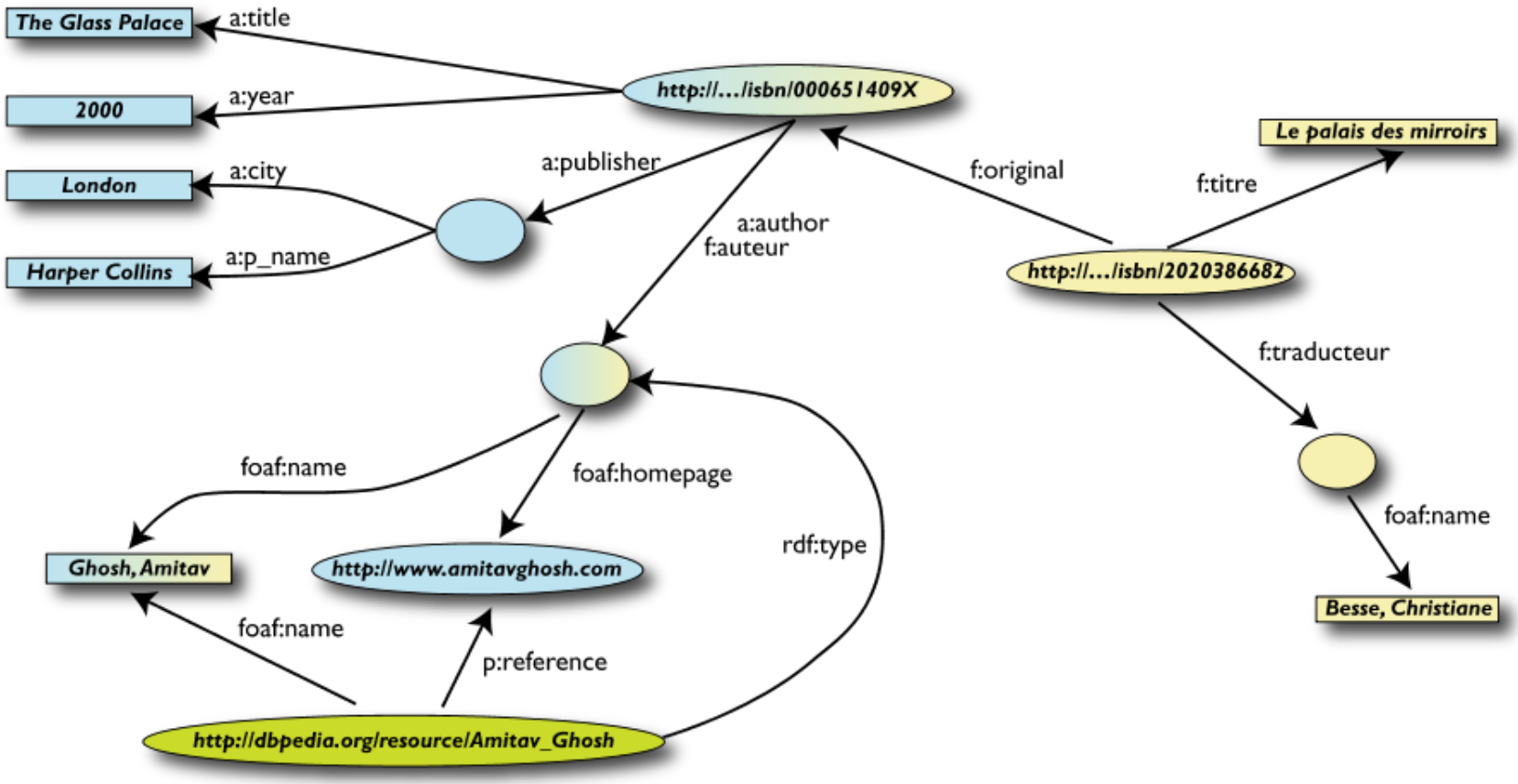




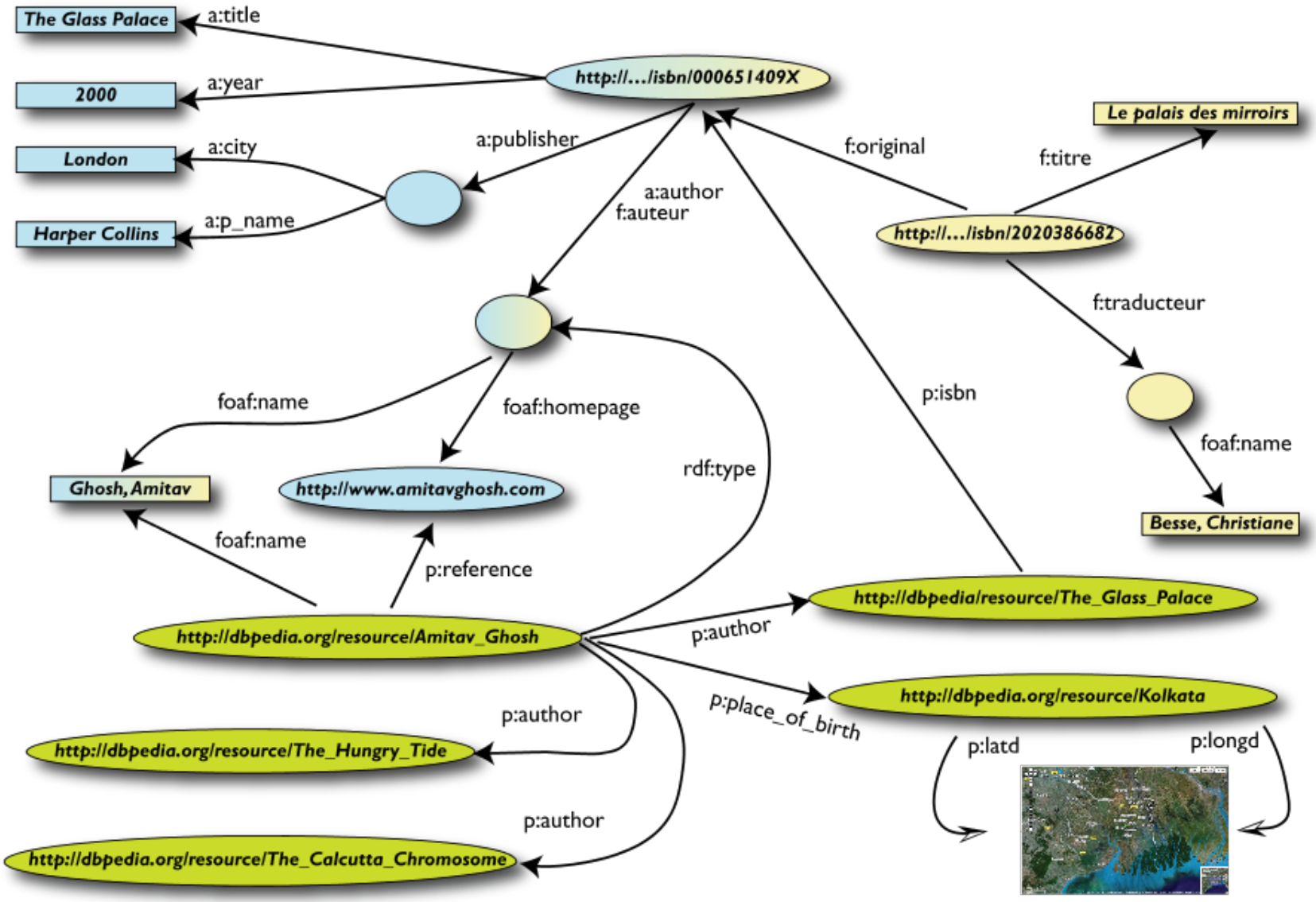
Start making richer queries!

- 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"

Merge with Wikipedia Data



Merge with more Wikipedia Data



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 ISBNs 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

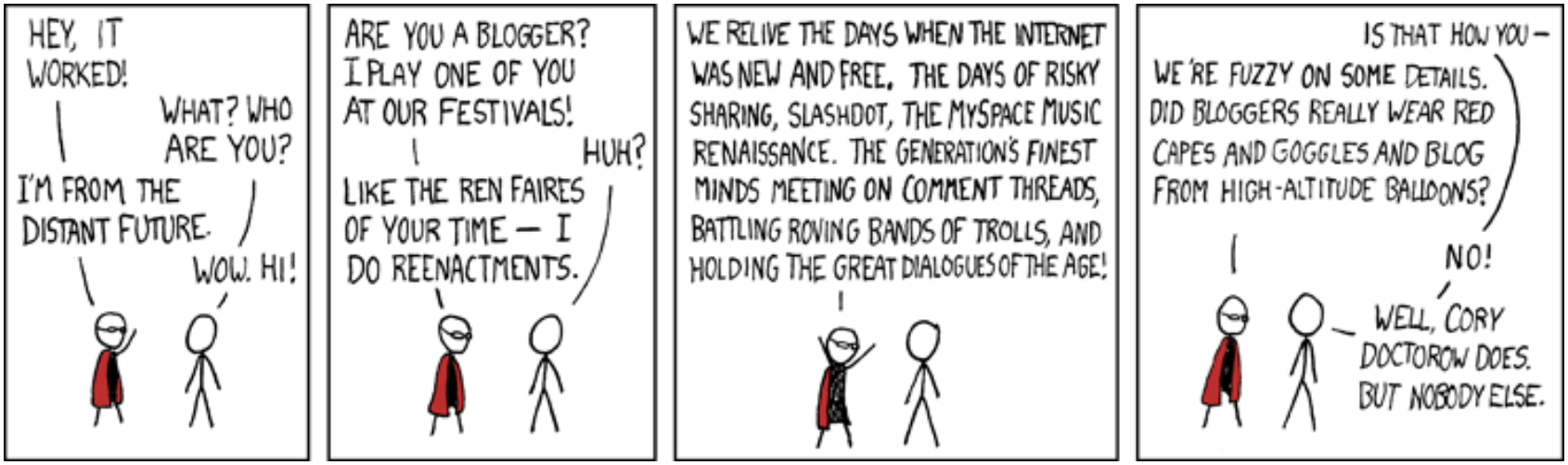
- We could add extra knowledge to the merged datasets
  - e.g., a full classification of various types of library data
  - geographical information
  - etc.
- This is where ontologies, extra rules, etc, come in
  - ontologies/rule sets can be relatively simple and small, or huge, or anything in between$\cdots$
- Even more powerful queries can be asked as a result!!!

Criticism of the Semantic Web

Doctorow's seven insurmountable obstacles to reliable metadata are:

1. People lie
2. People are lazy
3. People are stupid
4. Mission Impossible: know thyself
5. Schemas aren't neutral
6. Metrics influence results
7. There's more than one way to describe something

Cory Doctorow



A Canadian-British blogger, journalist, and science fiction author who serves as co-editor of the blog Boing Boing. He is an activist in favour of liberalising copyright laws and a proponent of the Creative Commons organization, using some of their licences for his books. Some common themes of his work include digital rights management, file sharing, and post-scarcity economics. (Wikipedia)

People lie

Metadata exists in a competitive world. Suppliers compete to sell their goods, cranks compete to convey their crackpot theories (mea culpa), artists compete for audience.

Thus:

- A search for any commonly referenced term at a search-engine like Altavista will often turn up at least one porn link in the first ten results.
- Your mailbox is full of spam with subject lines like "Re: The information you requested."
- Publisher's Clearing House sends out advertisements that holler "You may already be a winner!"
- Press-releases have gargantuan lists of empty buzzwords attached

People are lazy

Here in the Info-Ivory-Tower, we understand the importance of creating and maintaining excellent metadata for our information.

But info-civilians are remarkably cavalier about their information. Your clueless aunt sends you email with no subject line, half the pages on Geocities are called "Please title this page" and your boss stores all of his files on his desktop with helpful titles like "UNTITLED.DOC."

People are stupid

Even when there's a positive benefit to creating good metadata, people steadfastly refuse to exercise care and diligence in their metadata creation.

Take eBay: every seller there has a damned good reason for double-checking their listings for typos and misspellings. Try searching for "plam" on eBay. Right now, that turns up nine typoed listings for "Plam Pilots." Misspelled listings don't show up in correctly-spelled searches and hence garner fewer bids and lower sale-prices. You can almost always get a bargain on a Plam Pilot at eBay.

The fine (and gross) points of literacy – spelling, punctuation, grammar – elude the vast majority of the Internet's users. To believe that J. Random Users will suddenly and en masse learn to spell and punctuate – let alone accurately categorize their information according to whatever hierarchy they're supposed to be using – is self-delusion of the first water.

Mission: Impossible – know thyself

In meta-utopia, everyone engaged in the heady business of describing stuff carefully weighs the stuff in the balance and accurately divines the stuff's properties, noting those results.

Simple observation demonstrates the fallacy of this assumption. When Nielsen used log-books to gather information on the viewing habits of their sample families, the results were heavily skewed to Masterpiece Theater and Sesame Street. Replacing the journals with set-top boxes that reported what the set was actually tuned to showed what the average American family was really watching: light entertainment.

People are lousy observers of their own behaviors. Entire religions are formed with the goal of helping people understand themselves better; therapists rake in billions working for this very end.

Schemas aren't neutral

In a given sub-domain, say, Washing Machines, experts agree on sub-hierarchies, with classes for reliability, energy consumption, color, size, etc.

Nothing could be farther from the truth. Any hierarchy of ideas necessarily implies the importance of some axes over others. A manufacturer of small, environmentally conscious washing machines would draw a hierarchy that looks like this:

```
Energy consumption:
    Water consumption:
        Size:
            Capacity:
                Reliability
```

While a manufacturer of glitzy, feature-laden washing machines would want something like this:

```
Color:
    Size:
        Programmability:
            Reliability
```

The conceit that competing interests can come to easy accord on a common vocabulary totally ignores the power of organizing principles in a marketplace.

Metrics influence results

Ranking axes are mutually exclusive: software that scores high for security scores low for convenience, desserts that score high for decadence score low for healthiness. Every player in a metadata standards body wants to emphasize their high-scoring axes and de-emphasize (or, if possible, ignore altogether) their low-scoring axes.

It's wishful thinking to believe that a group of people competing to advance their agendas will be universally pleased with any hierarchy of knowledge. The best that we can hope for is a detente in which everyone is equally miserable.

There's more than one way to describe something

"No, I'm not watching cartoons! It's cultural anthropology."

"This isn't smut, it's art."

"It's not plagiarism, it's borrowing!"

Reasonable people can disagree forever on how to describe something. Arguably, your Self is the collection of associations and descriptors you ascribe to ideas. Requiring everyone to use the same vocabulary to describe their material denudes the cognitive landscape, enforces homogeneity in ideas.

And that's just not right.

Other Issues

Other reasons that result in metadata becoming obsolete are:

1. Data may become irrelevant in time

2. Data may not be updated with new insights

Reliable Metadata

- Information people use
  - Number of links into a page
  - Text on the page

- Even this gets gamed (link farms, spam pages, ...)

Semantic Web and NLP

- The Semantic Web is about structuring data
- Text Mining is about unstructured data
- There is much more unstructured than structured data
  - NLP can infer structure
  - NLP makes the Semantic Web feasible
  - the Semantic Web can be a resource for NLP

  (Computational) Linguistics is useful

Web 1.0/2.0/3.0

1.0 The read-only web (1993–2001)
interaction off-web (email, bulletin-boards, news-groups)

2.0 The read/write/share web (2001–2011)
social media

3.0 The personalized web (2011–)

The following pages are adapted from: CS Ramachandran (2011) "Spinning the Web –1.0, 2.0 or 3.0??" In *Blog: Random Thoughts — Thinking aloud about all things from Tech to life's bottlenecks* July 11, 2011 http://csramachandran.wordpress.com/2011/07/11/what-is-web3-0/ (accessed on 2011-03-26)

Web 1.0

- A read only web
- A nascent internet space with around 250,000 websites
- Catering to 45 million users
- The era of Geocities & Hotmail
- Dominated by Netscape and Internet Explorer browsers.
- It was all about read-only content and static HTML website
- Basic website design' s with Gif files
- Absence of dynamic user-generated content.

The era of Web 1.0 lasted from 1993 till the dot.com bubble burst in 2001.

Web 2.0

- The read/write/share web
- Rise of user-generated content
- Information sharing through blogs, social networks and image / video sites
- Enclyopedia Britannica replaced by Wikipedia
- Explosive growth of users –Over a billion users by 2006
- Brands and Websites start to focus on Social share and community buildings
- Emergence of Mobile Internet and Mobile Applications
- Rich User Experiences

Web 3.0

- The personalized web –iGoogle, My Yahoo, Personalized News Sites
- The ubiquitous web: convergence of the virtual and physical world
- The semantic Web –The web understands and anticipates requests
- The more you use the Web, the more your browser learns about you
- More localization of results –Google Places, Local Business Results
- Emergence of widgets and mashups (Combination of two applications –eg: review a place from Google Maps?)
- Behavioral targeting and advertising
- Evolution of a personal assistant who knows practically everything about you and your tastes
- More access points: computers, phones, cameras, TVs, ...

Readings/Resources

- The Semantic Web and Ontologies
  http://www.obitko.com/tutorials/ontologies-semantic-web/
- Criticism of the Semantic Web
  http://www.well.com/~doctorow/metacrap.htm
- XML: http://en.wikipedia.org/wiki/XML
- Semantic Web in general: semanticweb.org
- Nice presentation:
  http://www.w3.org/2004/Talks/0120-semweb-umich/.