The Semantic Web

DEFINITIONS & APPLICATIONS

POLITECNICO
DI TORINO

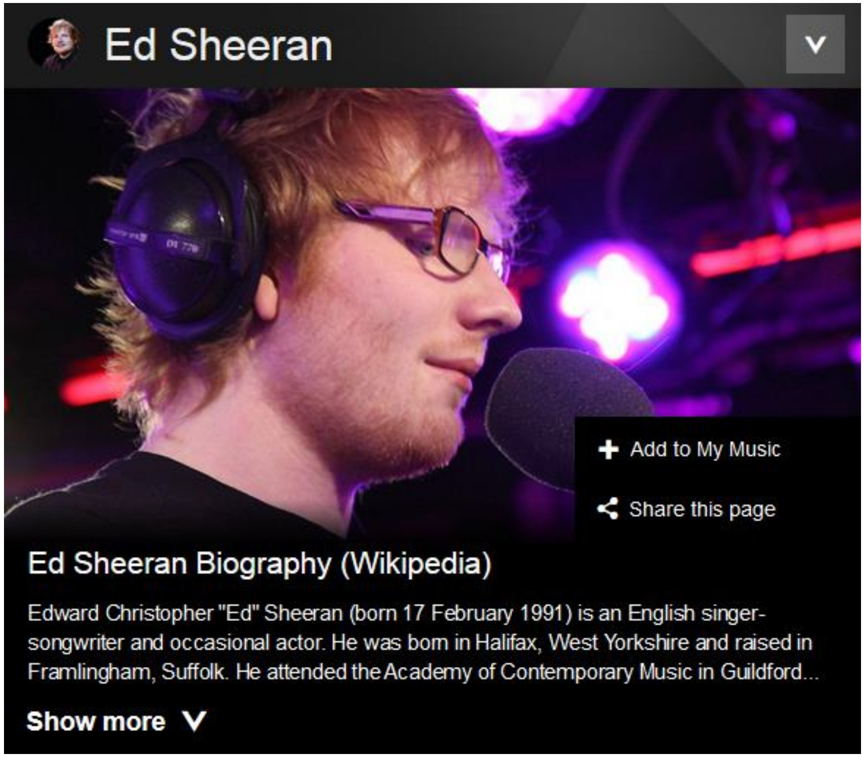
e-Lite

Data on the Web

- There are more an more data on the Web
  - Government data, health related data, general knowledge, company information, flight information, restaurants,…
  - This is evident!!!
- More and more applications rely on the availability of that data
  - Is that equally evident?
  - Let' s consider an example…

Example: how to build a music site (1)

- Site editors search the Web for new facts
  - May discover further links while searching
- They update the site manually
- And the site gets soon out-of-date



Ed Sheeran

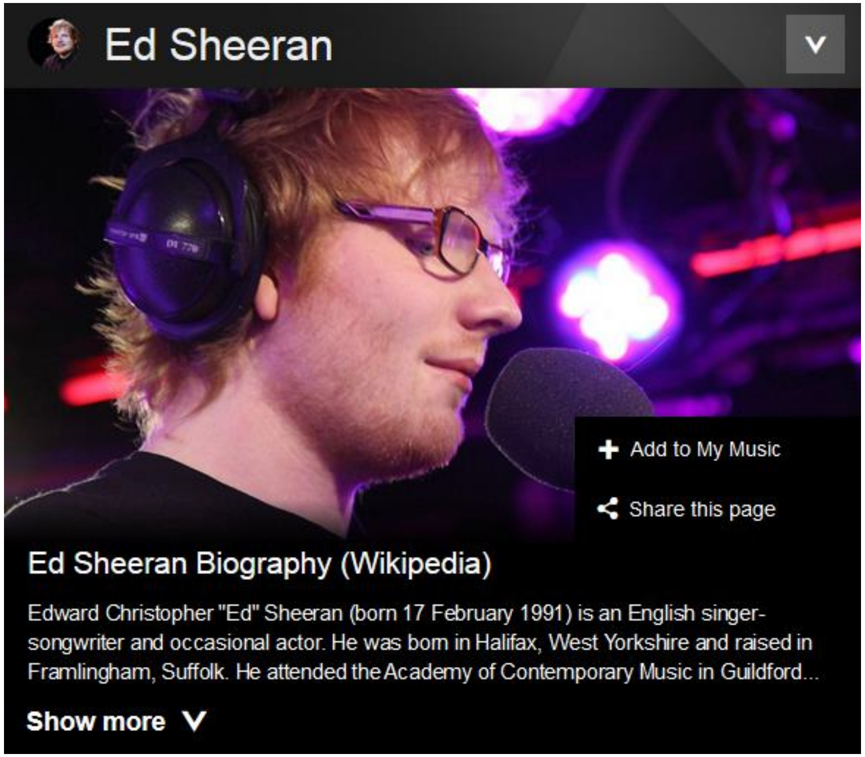
Ed Sheeran Biography (Wikipedia)

Edward Christopher "Ed" Sheeran (born 17 February 1991) is an English singer-songwriter and occasional actor. He was born in Halifax, West Yorkshire and raised in Framlingham, Suffolk. He attended the Academy of Contemporary Music in Guildford...

http://www.bbc.co.uk/music

Example: how to build a music site (2)

- Editors search the Web for new data published on Web sites
- They “scrape” the sites with a program to extract the information
    - i.e., write some code to incorporate the new data
- Easily get out of date again…

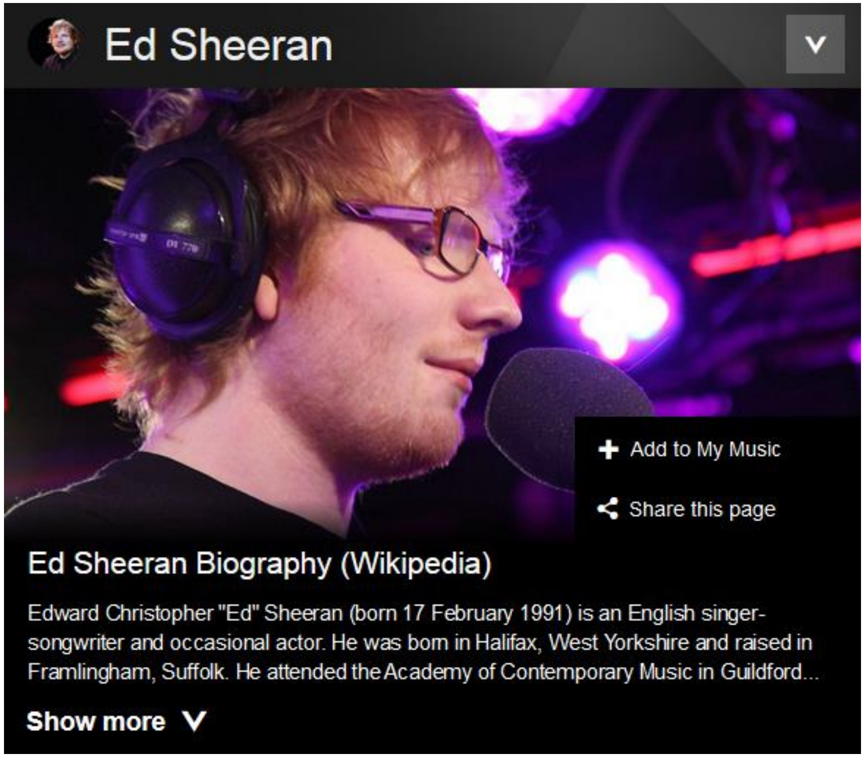


Ed Sheeran Biography (Wikipedia)

Edward Christopher "Ed" Sheeran (born 17 February 1991) is an English singer-songwriter and occasional actor. He was born in Halifax, West Yorkshire and raised in Framlingham, Suffolk. He attended the Academy of Contemporary Music in Guildford...

Example: how to build a music site (3)

- Editors search the Web for new data via APIs
- They understand …
  - input, output, arguments, datatypes, …
- They write some code to incorporate the new data
- Easily get out of date again…



Ed Sheeran

Ed Sheeran Biography (Wikipedia)

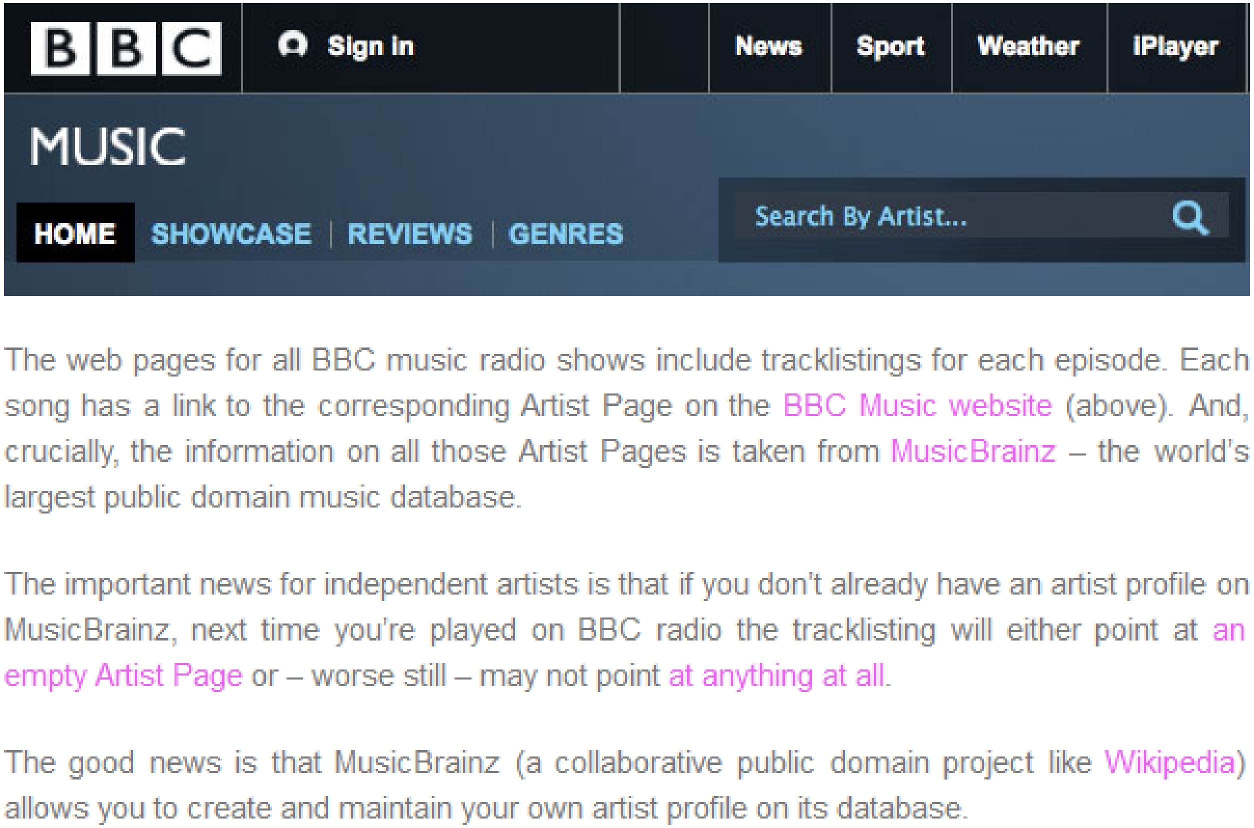
Edward Christopher "Ed" Sheeran (born 17 February 1991) is an English singer-songwriter and occasional actor. He was born in Halifax, West Yorkshire and raised in Framlingham, Suffolk. He attended the Academy of Contemporary Music in Guildford...

The choice of the BBC

https://www.w3.org/2001/sw/sweo/public/UseCases/BBC/

- The BBC is the largest broadcasting corporation in the world
- Use external, public datasets
  - Wikipedia, MusicBrainz, …
- They are available as data
  - data can be extracted using, e.g., HTTP requests or standard queries
- In short …
  - Use the Web of data as a content management system
  - Use the community at large as content editors

MusicBrainz: and why it matters



The web pages for all BBC music radio shows include tracklistings for each episode. Each song has a link to the corresponding Artist Page on the BBC Music website (above). And, crucially, the information on all those Artist Pages is taken from MusicBrainz – the world's largest public domain music database.

The important news for independent artists is that if you don't already have an artist profile on MusicBrainz, next time you're played on BBC radio the tracklisting will either point at an empty Artist Page or – worse still – may not point at anything at all.

The good news is that MusicBrainz (a collaborative public domain project like Wikipedia) allows you to create and maintain your own artist profile on its database.

http://freshonthenet.co.uk/musicbrainz/

Key benefits of using Semantic Web technology (according to BBC)

- Usability: making a site around the things people care and think about
- User Experience: having meaningful predicates and granular, addressable resources, so that those resources can be visualized in new ways
- User Journeys: allowing users to make their own journeys across our content
  - On the BBC /nature, users can start making their own documentaries: they can start on an animal, watch a programme clip, follow a link to a related habitat, read about that habitat and so on…
- One page per thing: making our resources part of the Web and therefore linkable and discoverable
- Our web site is our API: one URI for both machines and web browsers
  - Our web site can be used by third parties to create new products, e.g., URIPlay, TestTubeTelly, FanHubz or Channelography
- Loosely coupled development: different teams can work together in a loosely coupled fashion; each team focuses on their domain of interest

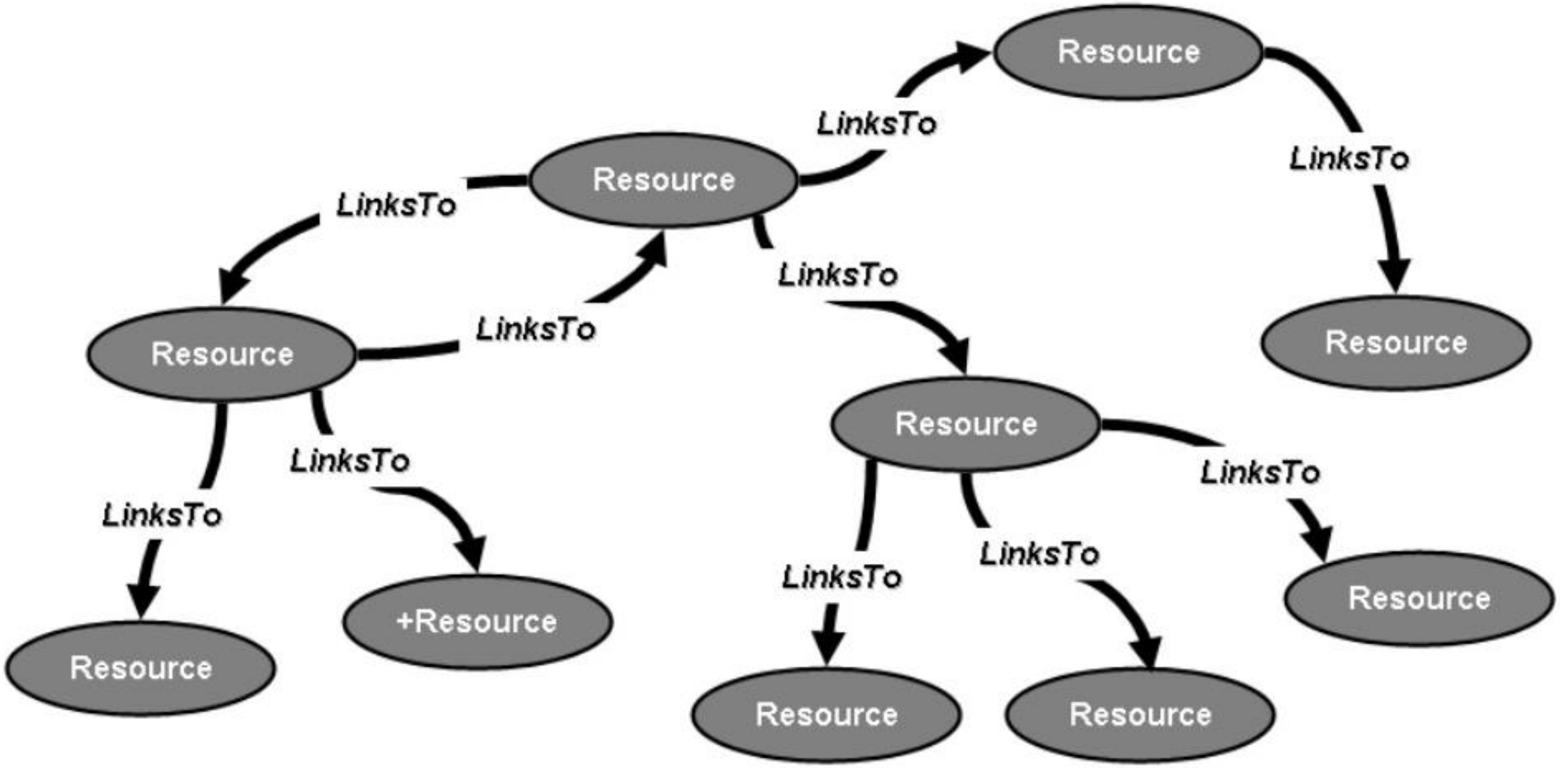
Data on the Web

- We need a proper infrastructure for a real Web of data
  - Data is available on the Web, and accessible via standard Web technologies
  - Data are interlinked over the Web: i.e., data can be integrated over the Web
- This is the role of the Semantic Web technologies

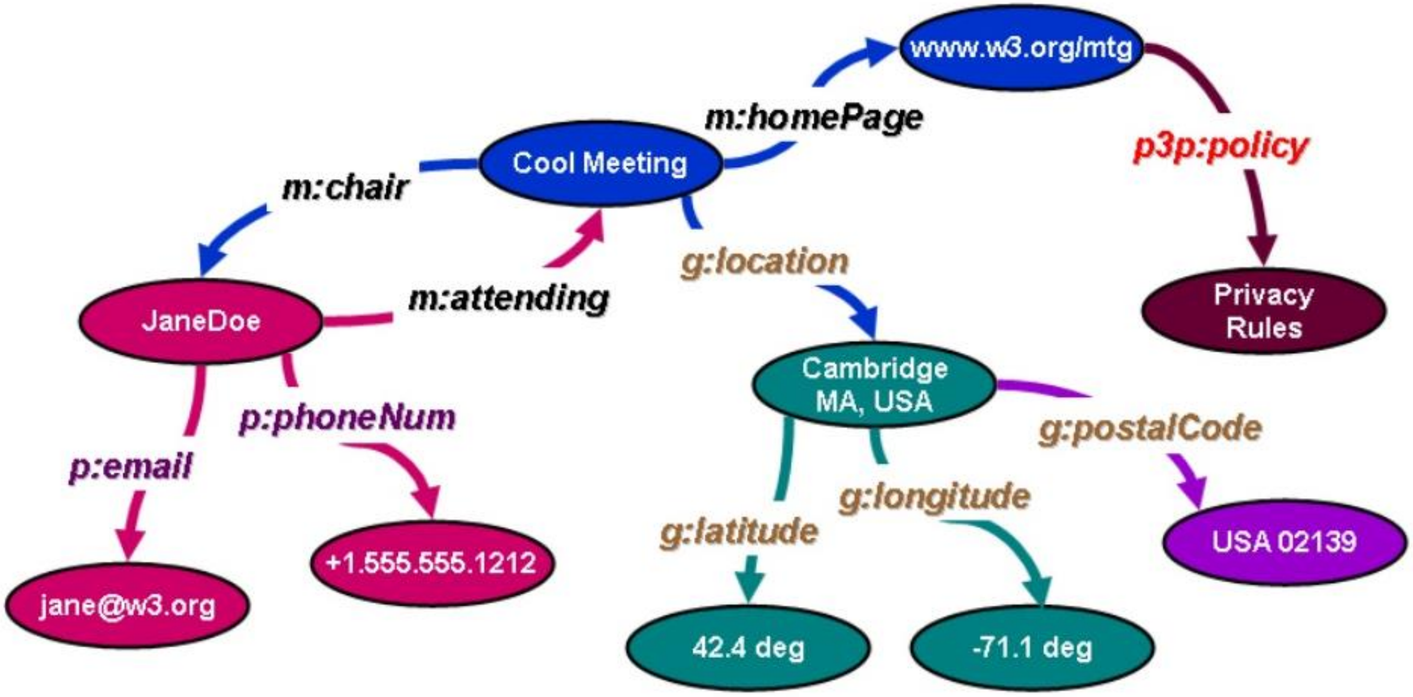
Definition

- The Semantic Web is a Web of linked data
  - dates and titles and numbers and chemical properties and any other data one might conceive of
- The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network
  - Web information must be machine-readable
- Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data

The Web is about documents

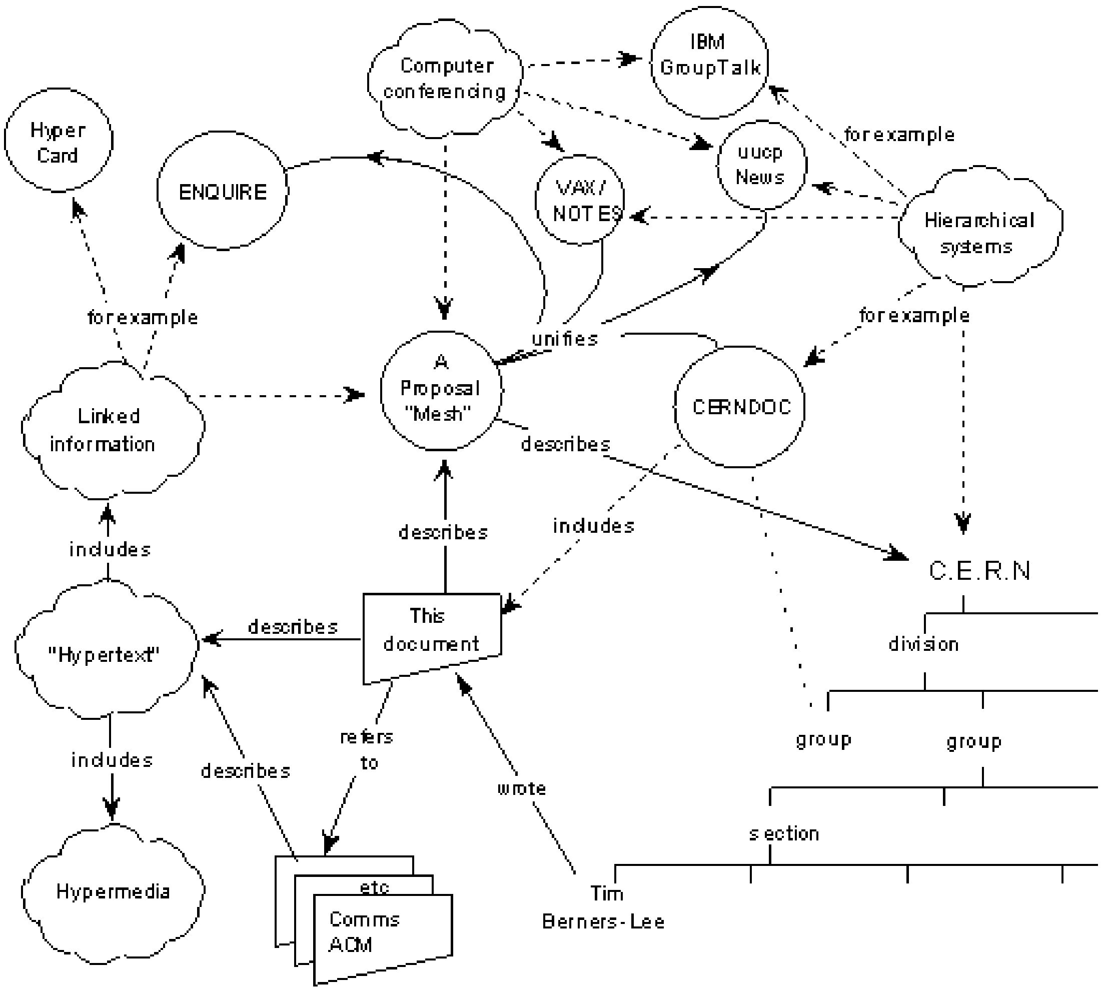


The Semantic Web is about “things”



A curiosity

- The original Web concept (1989)



What is the Semantic Web?

- It’s a collection of standard technologies to realize a Web of Data
- It looks simple, but the devil is in the details
  - A common model has to be provided for machines to describe, query, …, the data and their connections
  - The “classification” of the terms can become very complex for specific knowledge areas: this is where ontologies, thesauri, …, enter the game

W3C GRDDL

W3C SKOS

W3C OWL

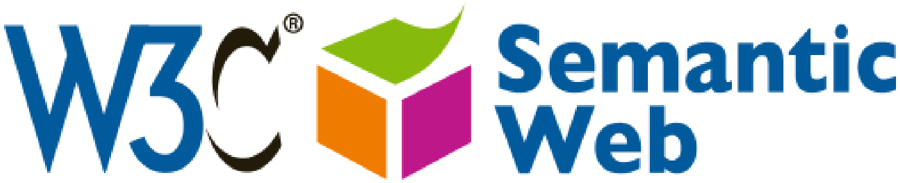
W3C POWDER

W3C SPARQL

W3C RDFa

W3C RDF

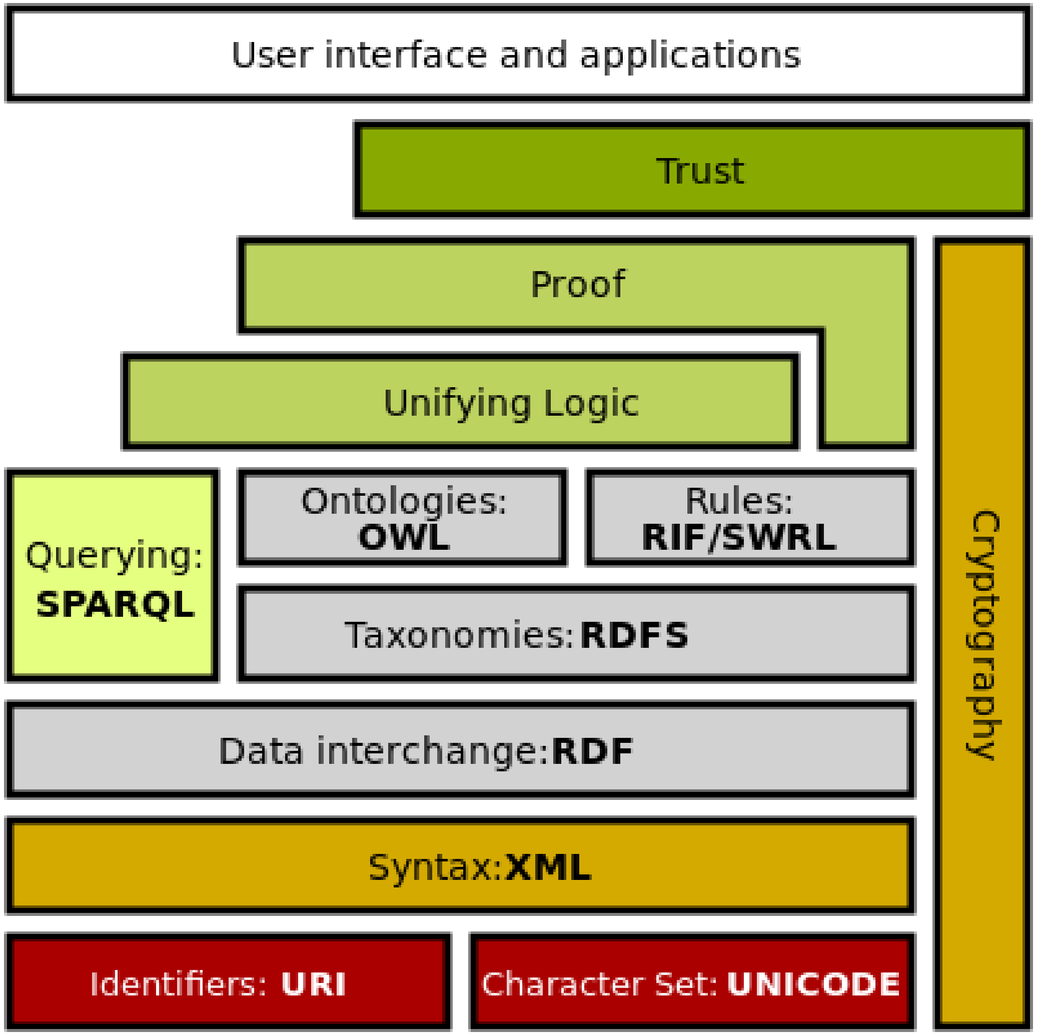
The W3C logo



- The three sides of the tri-color cube in the logo evoke the triplet of the RDF model
- The peeled back lid invites you to Open Your Data to the Semantic Web!

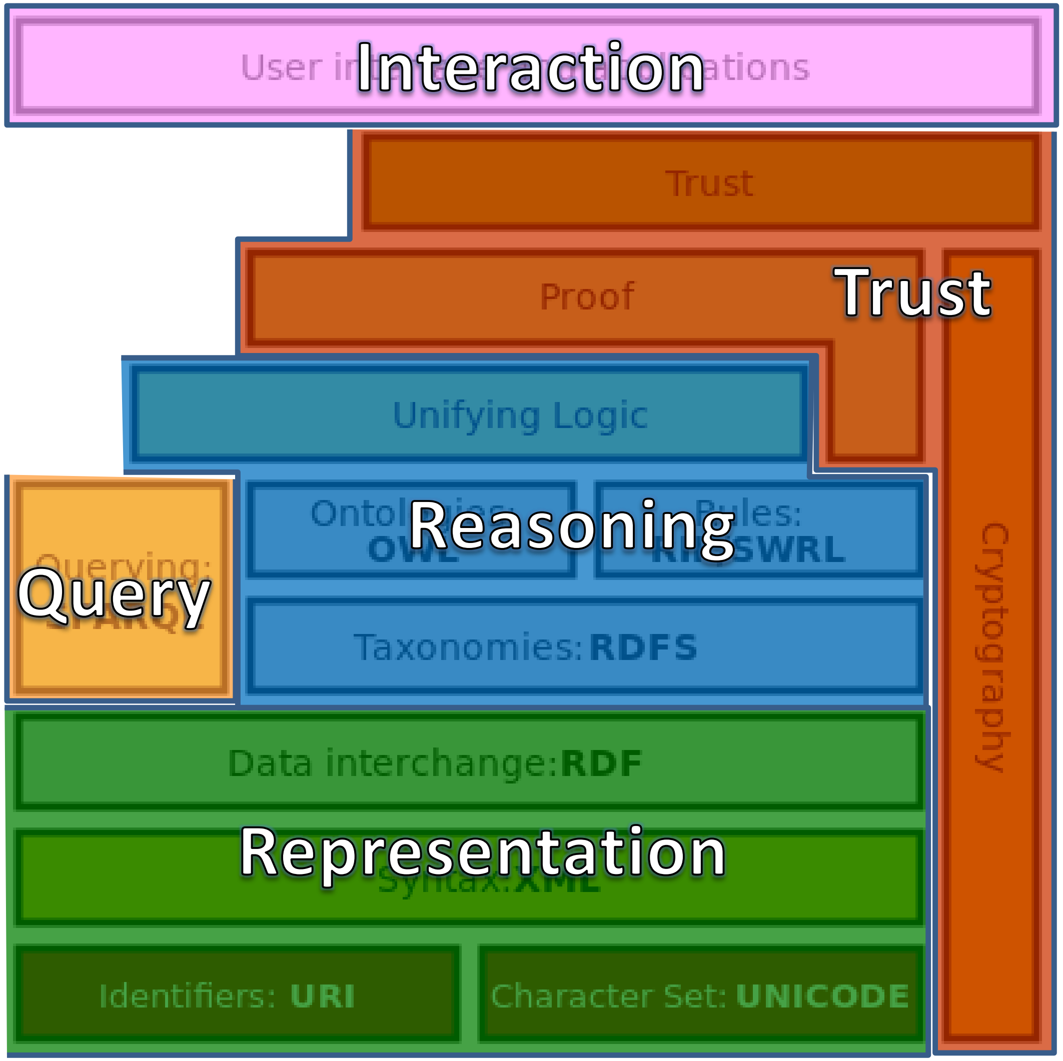
Semantic Web components

- The Semantic Web standard stack



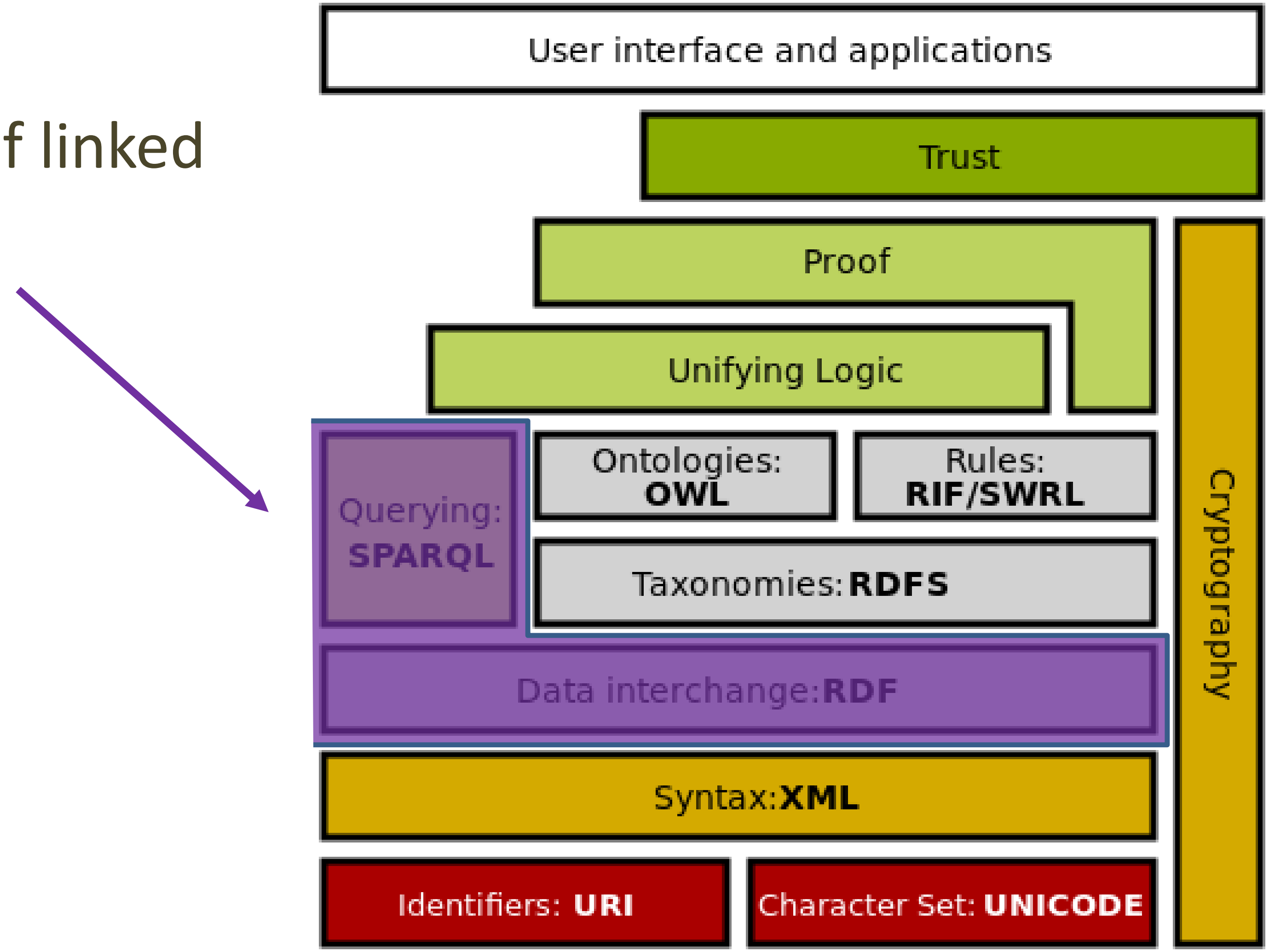
Semantic Web components

- We don't have yet standard solutions for trust



Semantic Web components

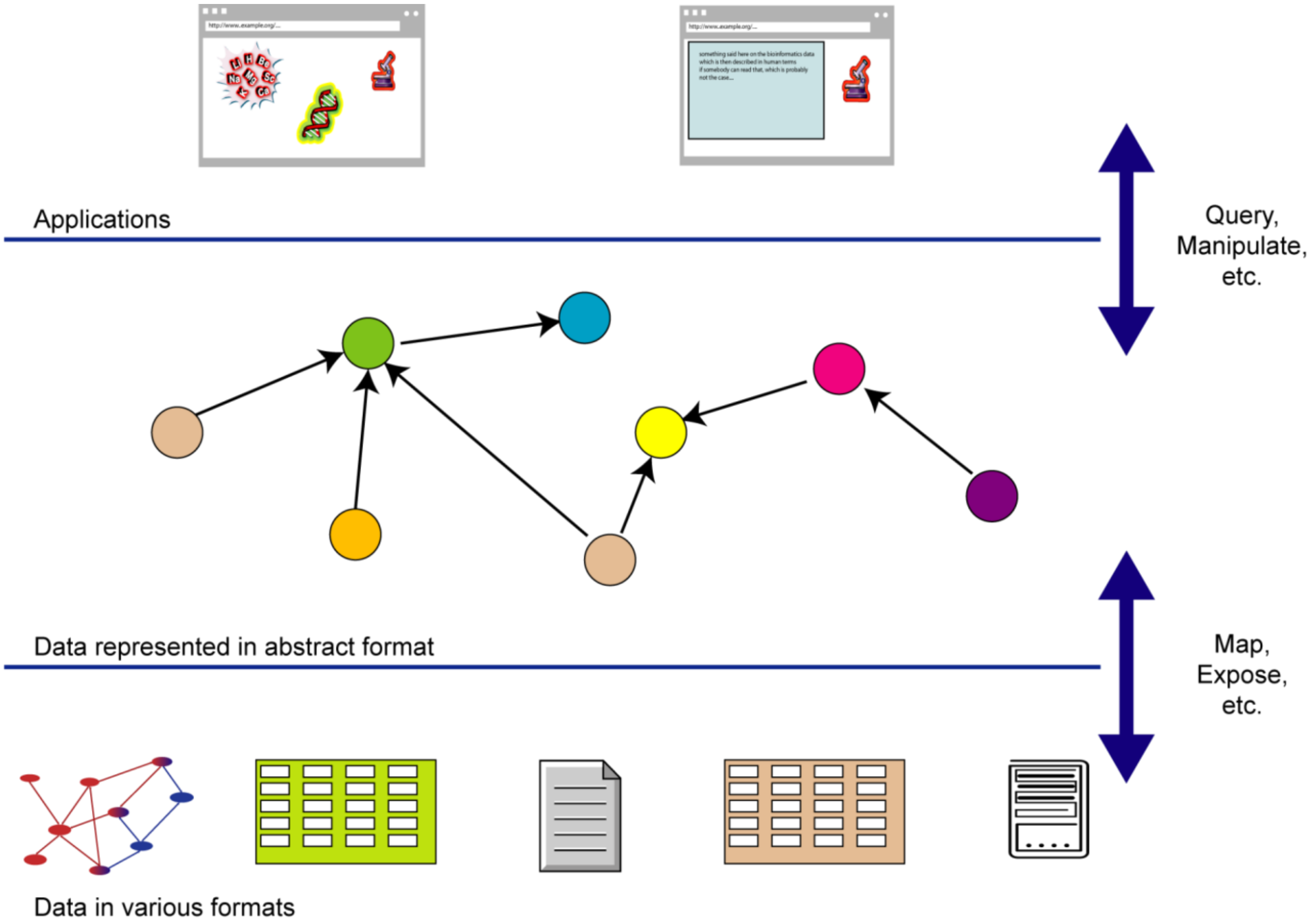
- A Web of linked data



To summarize… Semantic Web is

- A common set of technologies
  - ...enables diverse uses
  - ...encourages interoperability
- A coherent set of technologies
  - …encourage incremental application
  - …provide a substantial base for innovation
- A standard set of technologies
  - ...reduces proprietary vendor lock-in
  - ...encourages many choices for tool sets

What do Semantic Web solutions look like?



Applications

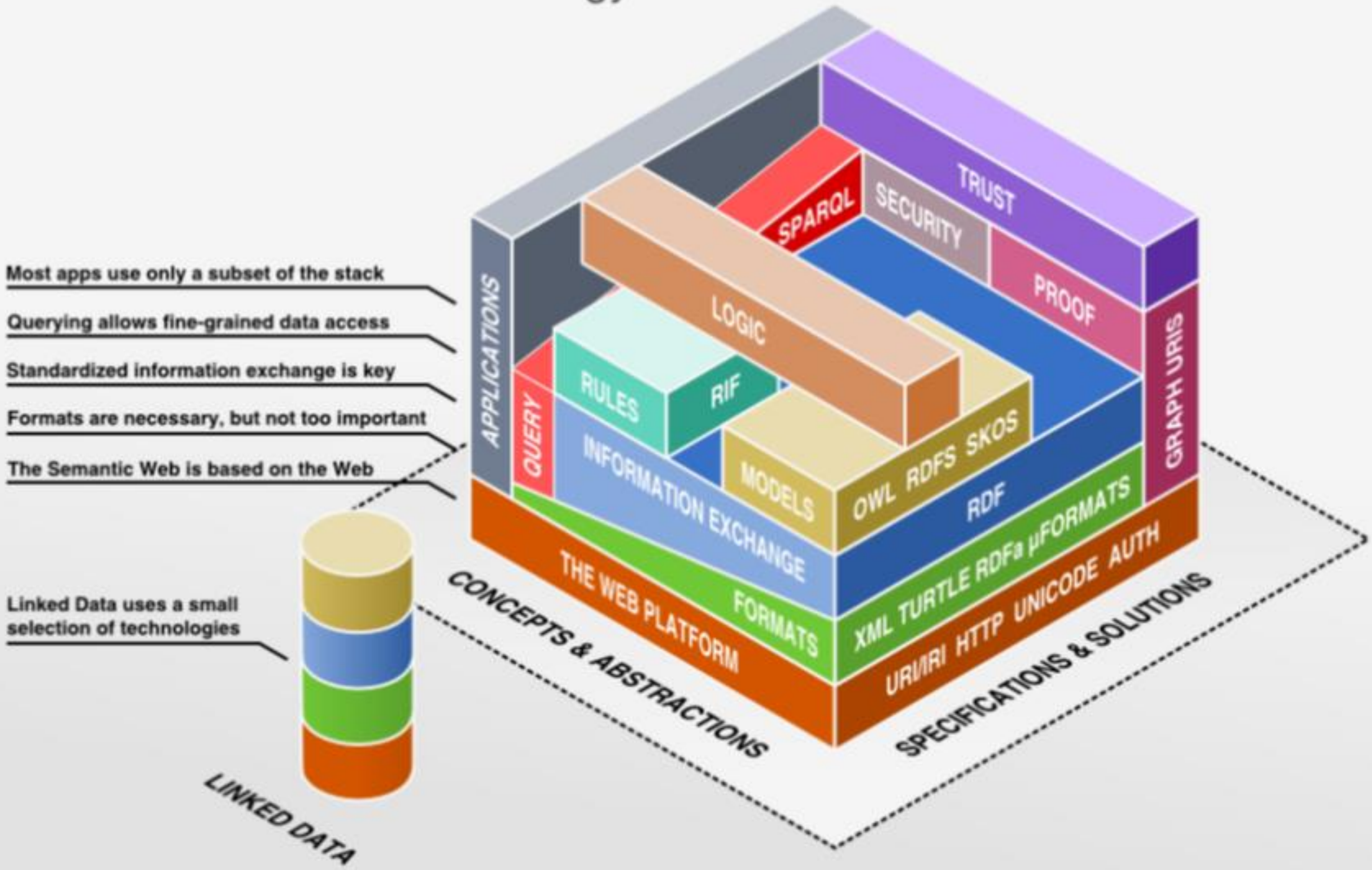
Query,
Manipulate,
etc.

Data represented in abstract format

Map,
Expose,
etc.

Data in various formats

The Semantic Web Technology Stack



Most apps use only a subset of the stack

Querying allows fine-grained data access

Standardized information exchange is key

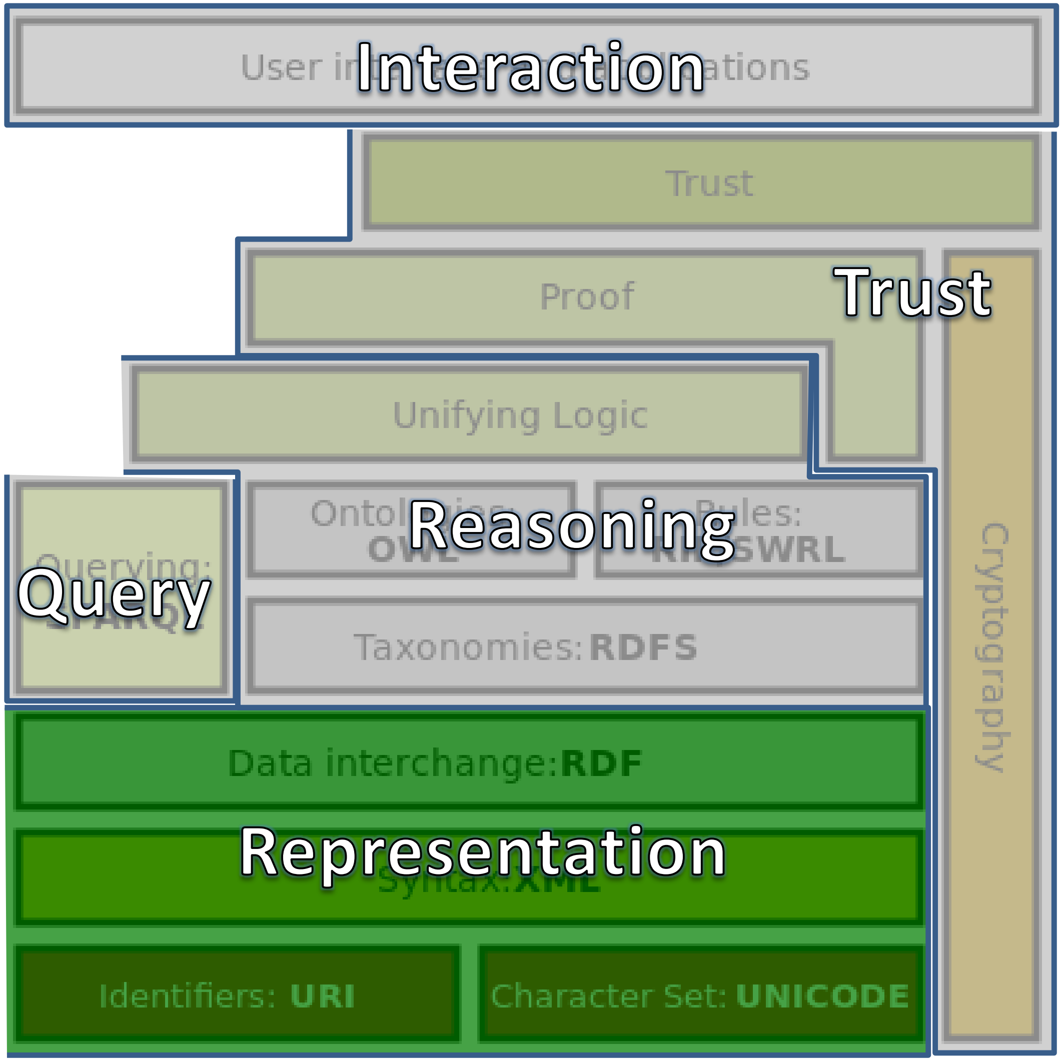
Formats are necessary, but not too important

The Semantic Web is based on the Web

Linked Data uses a small selection of technologies

Step 1: Representation

- The Semantic Web will enable machines to comprehend semantic documents and data, NOT human speech and writing



Metadata

- The Semantic Web foundation

THE TREE
THE DOOR
THE HOUSE
GARB CAN
SHIRT
PANTS
THE CAT
THE DOG

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

Resource and description

The title of this resource is "Introduction to the Semantic Web"

This resource was created on January 16th, 2017

The author of this resource is L. Farinetti

This resource is suitable for PhD students

This resource is related to computer science, knowledge representation and metadata

Resource

- Resource
  - Content, format, …
  - Access method dependent on format (I can read it if I "know" its language)
- Standardization (i.e. common language for applications) ???
  - Practically impossible …
  - Huge amount of existing information
  - Hundreds of human languages
  - Hundreds of computer languages (other word for formats)

Description

- Resource description
  - Independent of the format (I can read “people’s comments” about the resource… provided that I know the language in which the comment is written)
- Standardization (i.e. common language for applications) ???
  - Feasible
  - Smaller amount of information, possibly new
  - Solution: define a standard language for writing comments (“metadata” in semantic web terminology)

Resource and description

The title of this resource is "Introduction to the Semantic Web"

This resource was created on January 16th, 2017

Metadata

Field name = field value

The author of this resource is L. Farinetti

This resource is suitable for PhD students

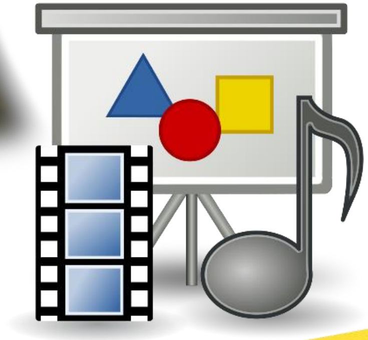
This resource is related to computer science, knowledge representation and metadata

Resource and description

Title = "Introduction to the Semantic Web"

Date = 2017-01-16

Resource



Author = L. Farinetti

Audience = PhD students

Topic = {computer science, knowledge representation, metadata}

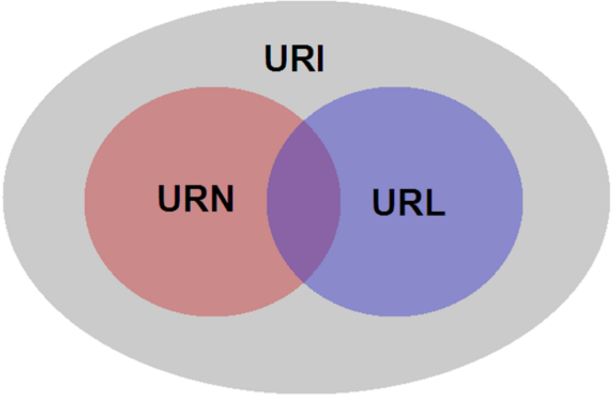
Meaningful metadata annotations

- Common language for describing resources
  - Resource description standards
- Common language for describing field names
  - Metadata standards
- Common language for describing field values
  - Metadata standards + controlled vocabularies
- Semantically rich descriptions to support reasoning
  - Knowledge representation techniques, ontologies

Common language for describing resources

- Resource Description Framework (RDF)
  - Resource = URI (retrievable, or not)
  - RDF is structured in statements
- A statement is a triple
  - Subject – predicate – object
  - Subject: a resource
  - Predicate: a verb / property / relationship
  - Object: a resource, or a literal string
- RDF has several syntaxes (Turtle, N3, …) and XML is one of those, known as RDF/XML
  - XML is a syntax while RDF is a data model

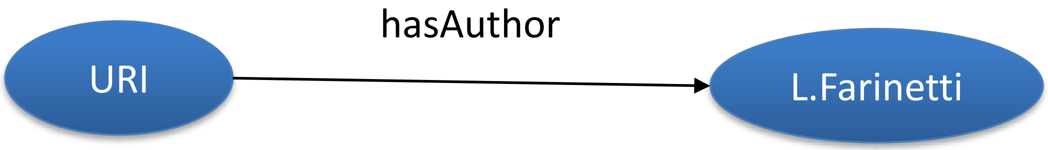
URIs: Uniform Resource Identifiers



- A URI provides a simple and extensible mean for identifying a resource
- A URI can be further classified as a locator (URL), a name (URN), or both
- A URL is a URI that, in addition to identifying a web resource, specifies the means of acting upon or obtaining the representation, specifying both its primary access mechanism and network location
- A URN is a URI that identifies a resource by name in a particular namespace
  - A URN can be used to talk about a resource without implying its location or how to access it

Common language for describing resources

- Diagram



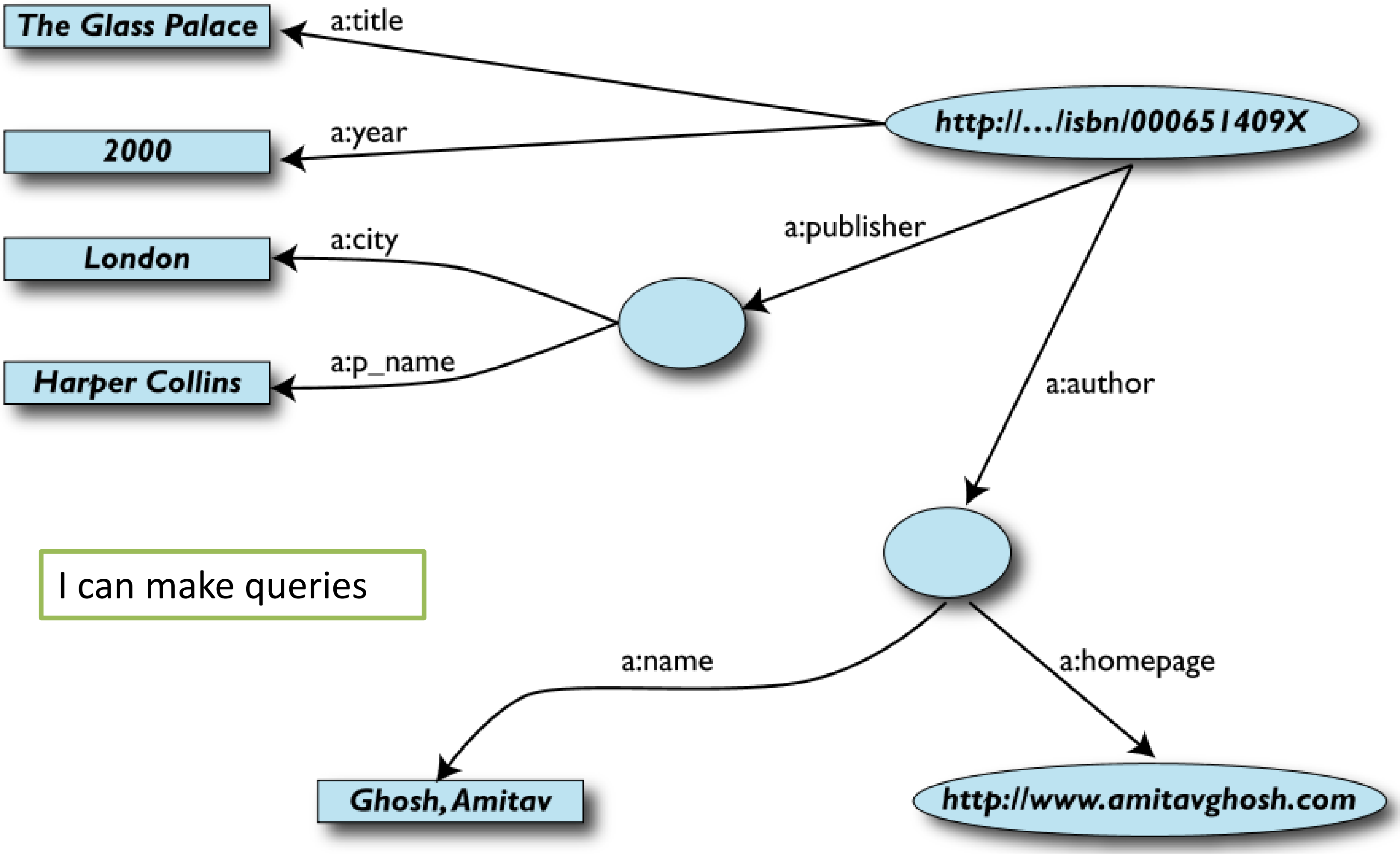
- Simple RDF assertion (triple)

```
triple (hasAuthor, URI, L.Farinetti)
```

- RDF in XML syntax

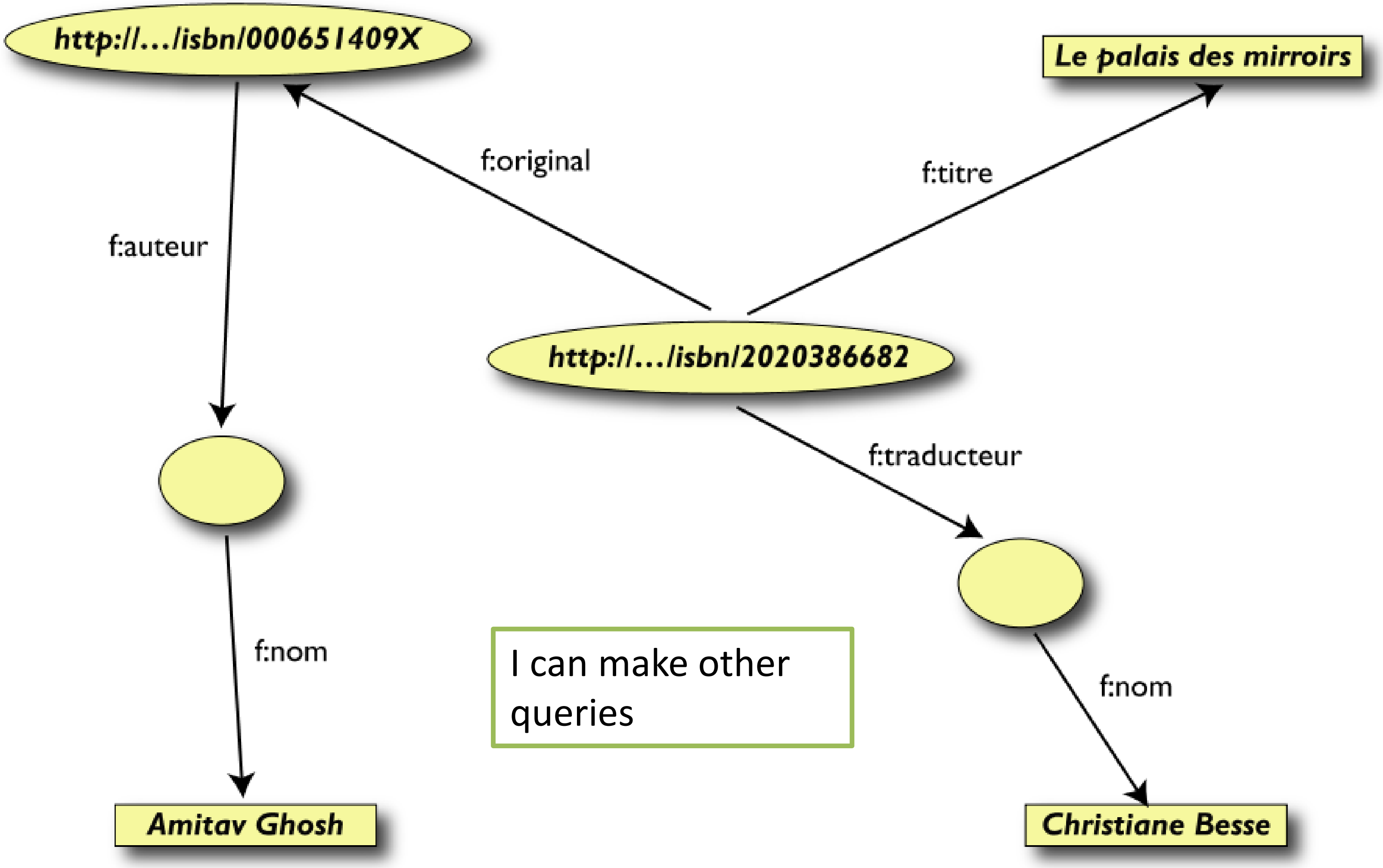
```
<RDF  xmlns=“http://www.w3.org/TR/ … ” >
  <Description about=“http://www.polito.it/semweb/intro”>
       <Author>L.Farinetti</Author>
  </Description>
</RDF>
```

A RDF example (1): some statements



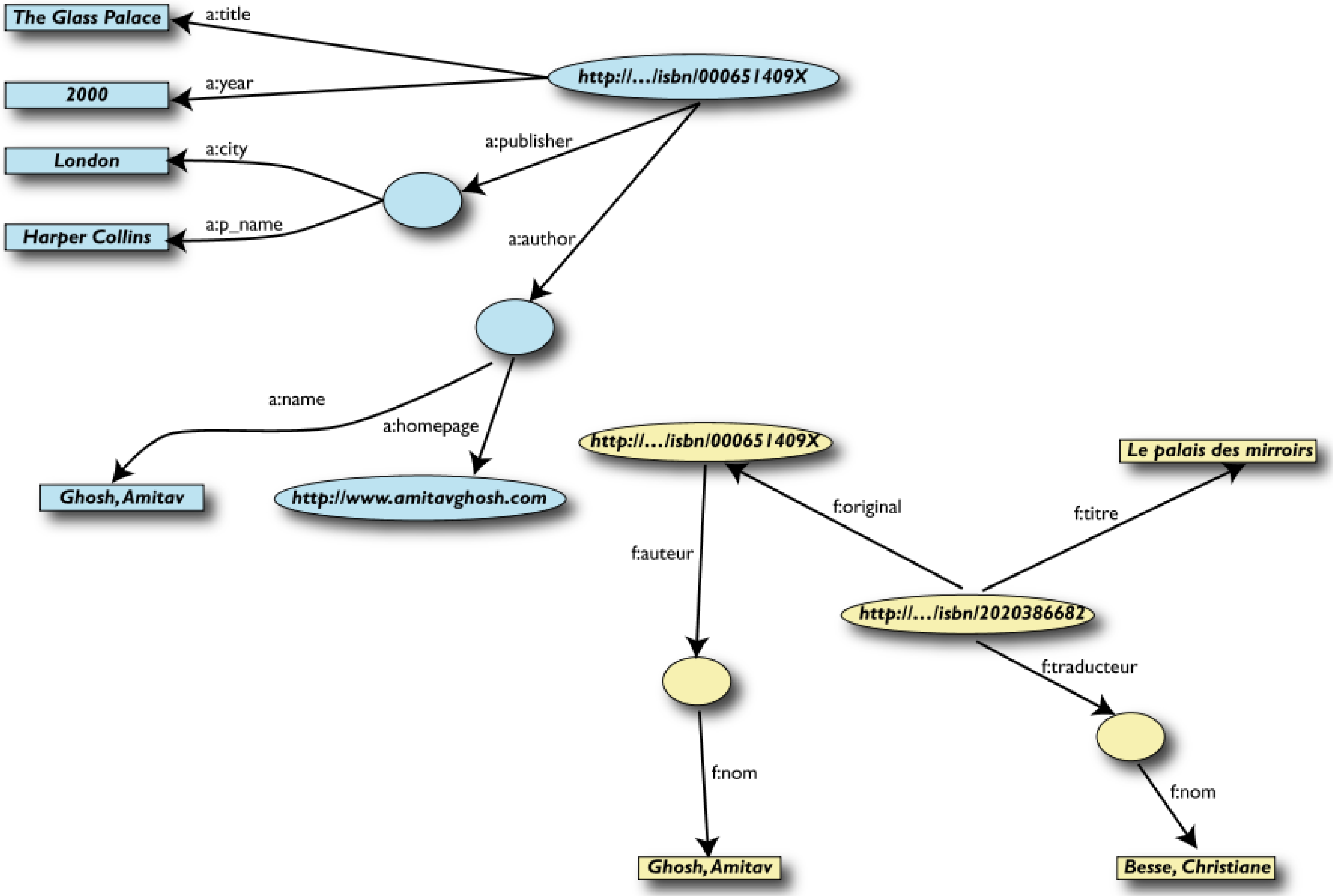
I can make queries

A RDF example (2): other statements

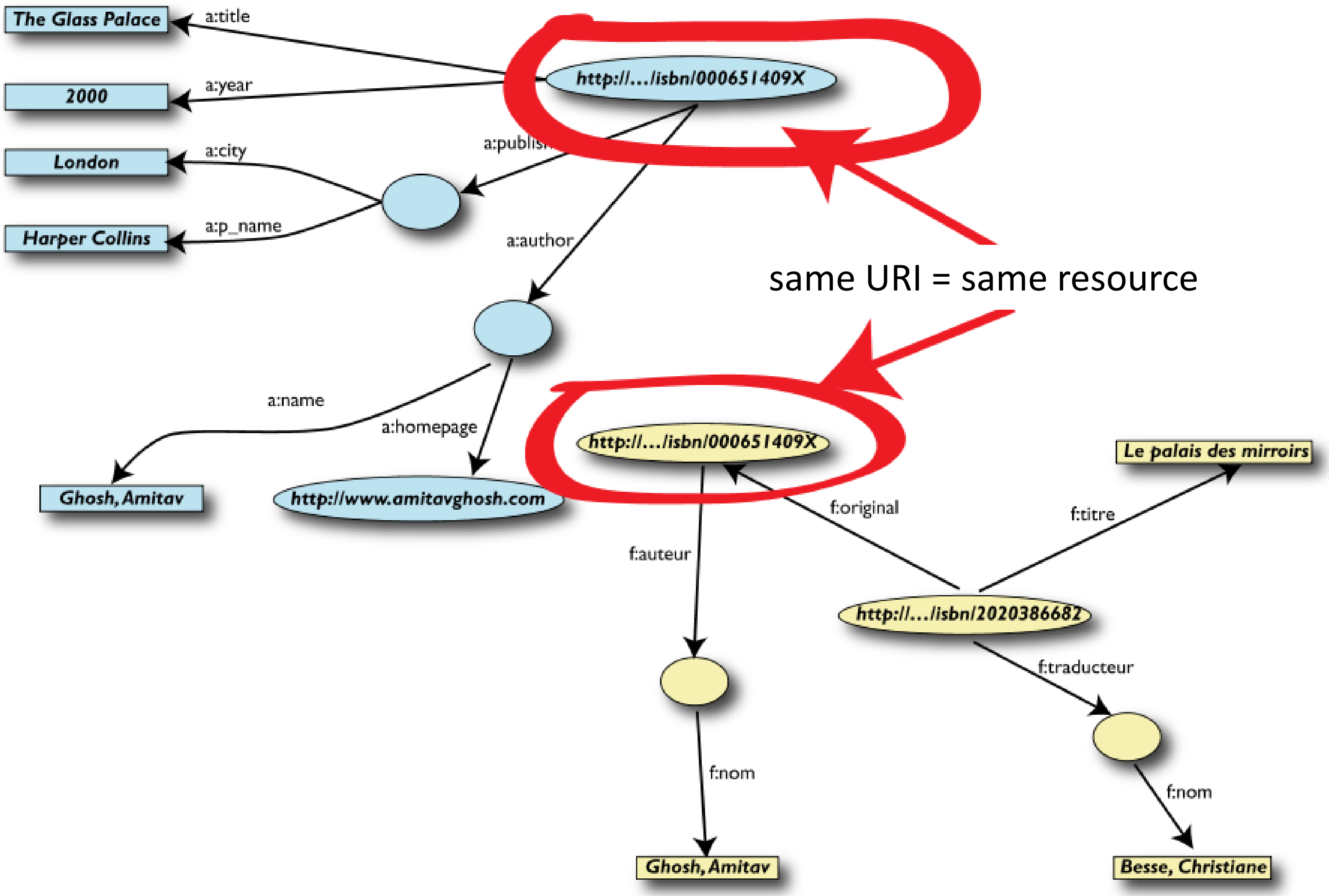


I can make other queries

A RDF example (3): same book!

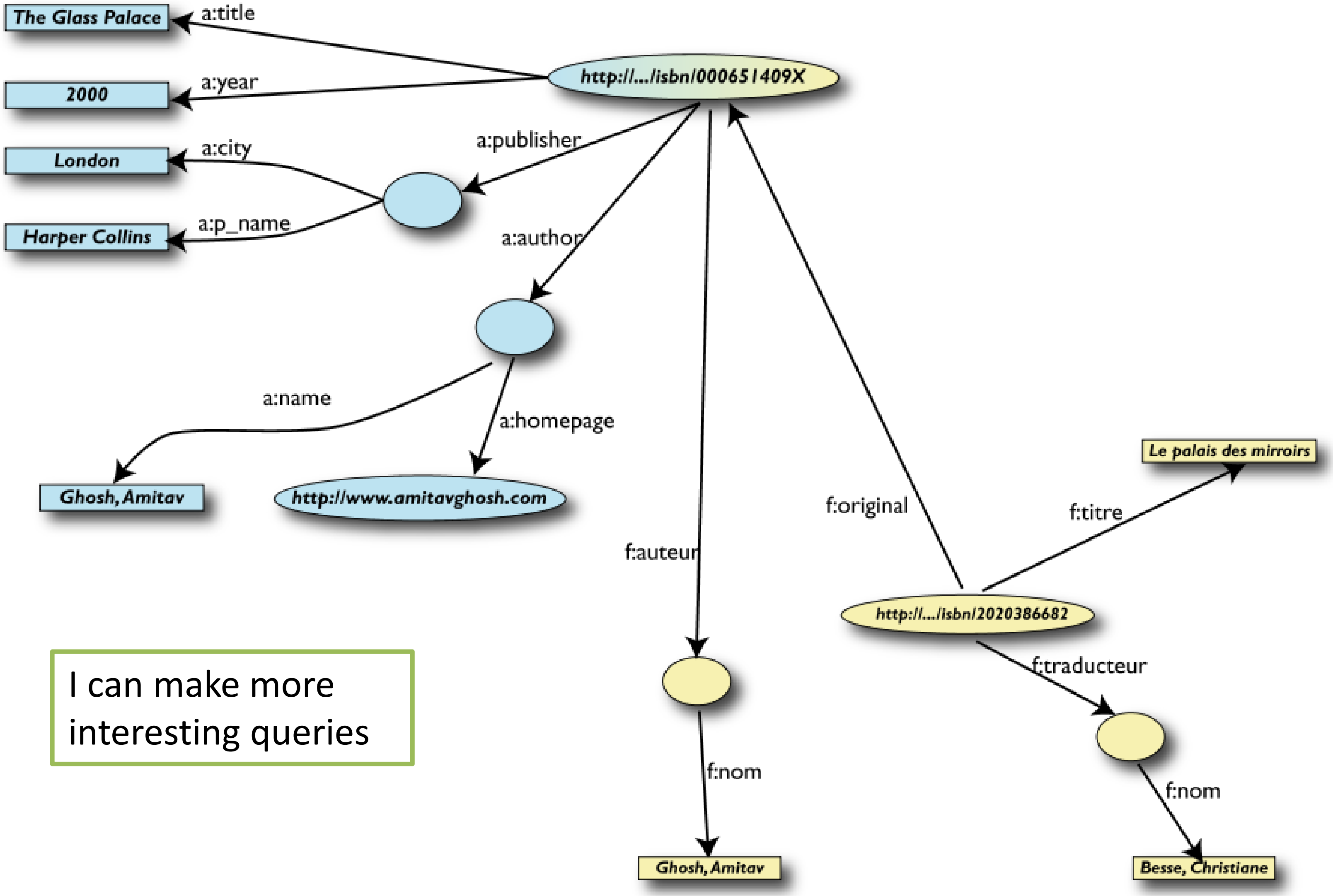


A RDF example (4): same URI



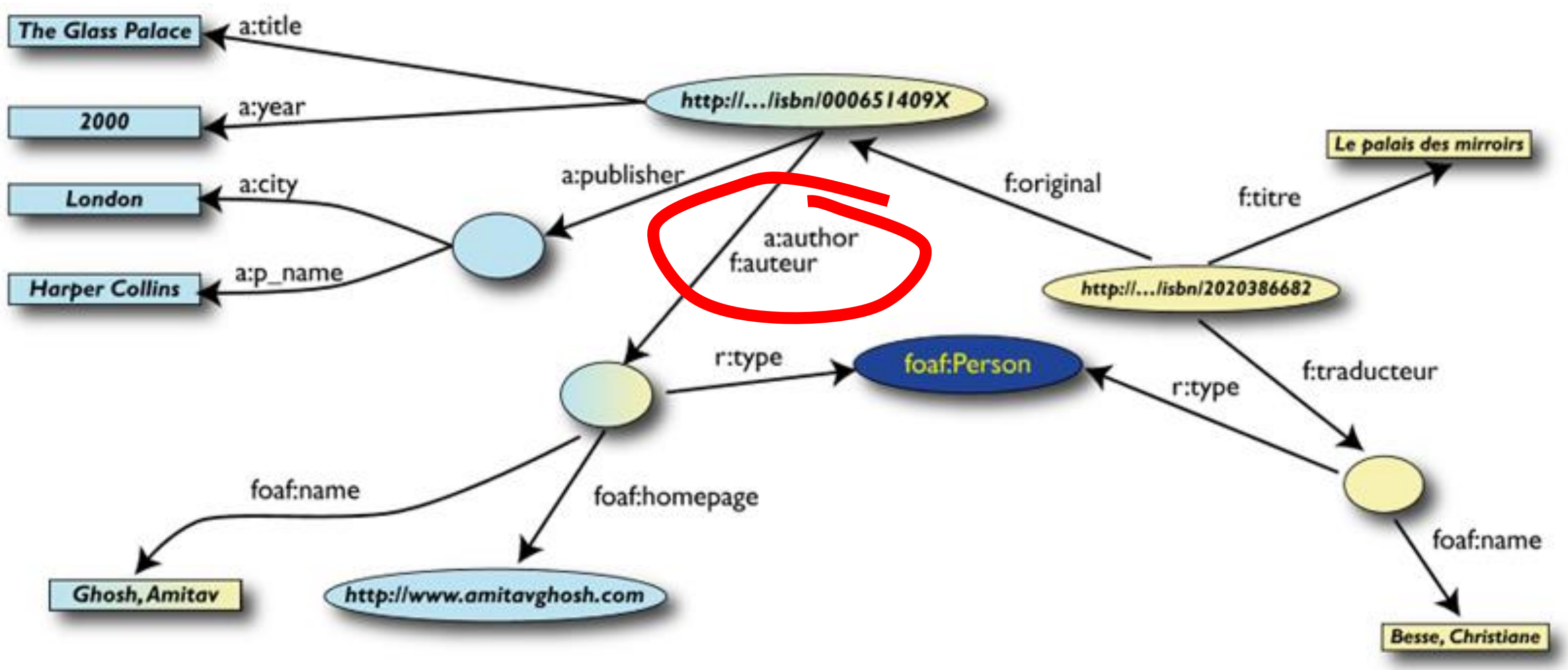
same URI = same resource

A RDF example (5): merge



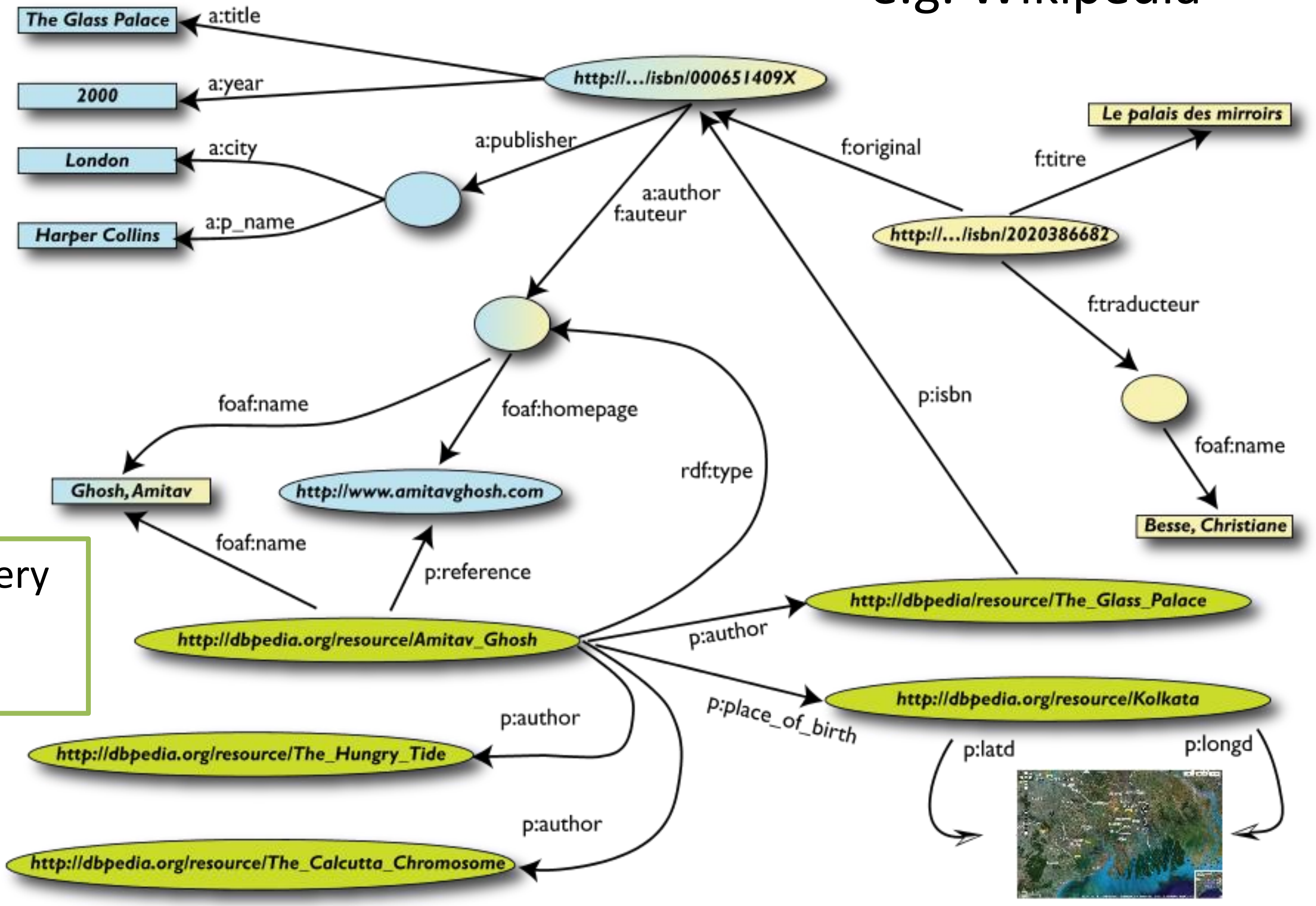
I can make more interesting queries

A RDF example (6): use extra knowledge



A RDF example (7): combine with different dataset

e.g. Wikipedia



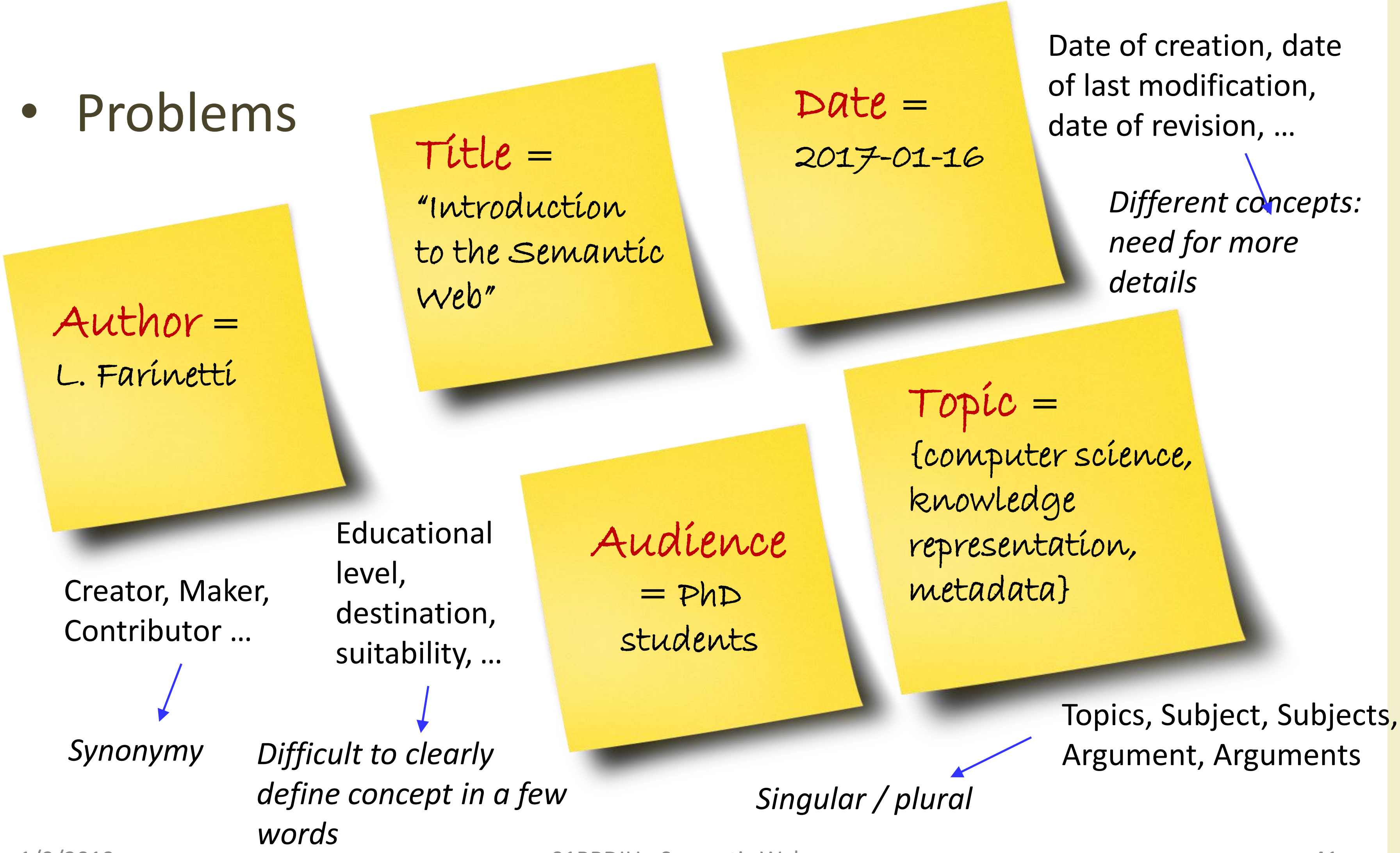
I can make very interesting queries

A RDF example (8): add more “power”

- We could add extra knowledge to the merged datasets
  - e.g., a full classification of various types of library data
  - geographical information
  - …
- This is where ontologies, extra rules, …, 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

Common language for field names

- Problems



Author = L. Farinetti

Creator, Maker, Contributor …

Synonymy

Title = "Introduction to the Semantic Web"

Date = 2017-01-16

Date of creation, date of last modification, date of revision, …

Different concepts: need for more details

Audience = PhD students

Educational level, destination, suitability, …

Difficult to clearly define concept in a few words

Topic = {computer science, knowledge representation, metadata}

Topics, Subject, Subjects, Argument, Arguments

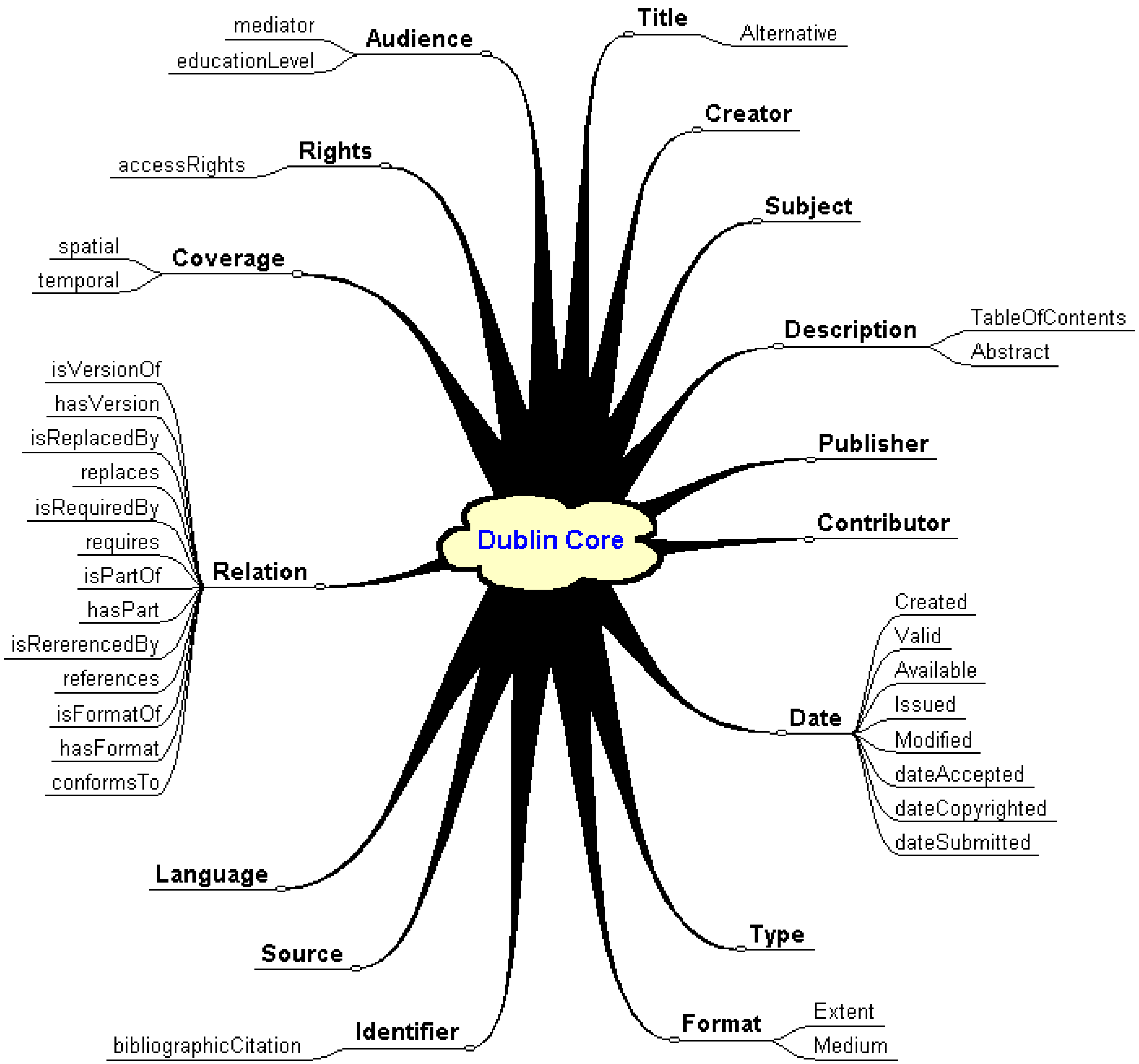
Singular / plural

Common language for field names

- Solution: metadata standards
- Many standardization bodies are involved
- Standards may be general …
  - e.g. Dublin Core (DC)
- … or may depend on goal, context, domain, …
  - e. g. educational resources (IEEE LOM), multimedia resources (MPEG-7), images (VRA), people (FOAF, IEEE PAPI), geospatial resources (GSDGM), bibliographical resources (MARC, OAI), cultural heritage resources (CIDOC CRM)

Example: Dublin Core

Dublin Core Metadata Initiative®
Making it easier to find information.



Common language for field values

- Problems
  - Value type

Title =
"Introduction to the Semantic Web"

type = string

Date =
2017-01-16

type = date

Author =
L. Farinetti

type = string
"standard" format?
Laura Farinetti, Farinetti
Laura, Farinetti L., …

Common language for field values

- Problems
  - Value type
  - Value restrictions? Freedom vs shared understanding

Audience

= PhD students

any value?
list of possible values?

Topic =
{computer science, knowledge representation, metadata}

any value?
any number of values?

Quality

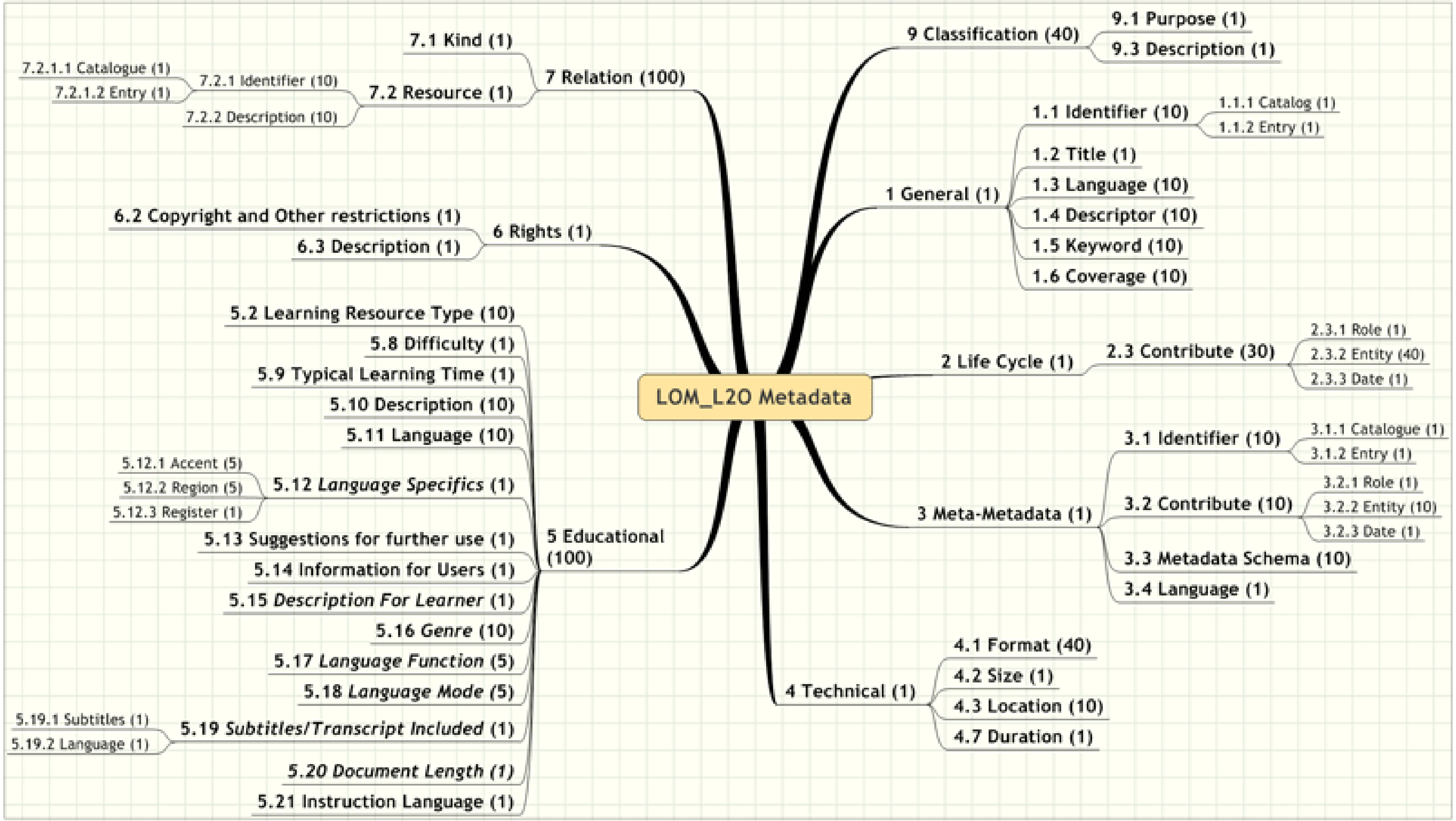
= high

High, medium, low?
1 to 5?
any value?

Common language for field values

- Solution: metadata standards + controlled vocabularies
- Metadata standards
  - Only some, and partially
- Controlled vocabularies
  - Explicit list of possible values

Example: IEEE LOM



LOM_L2O Metadata

1 General (1)
- 1.1 Identifier (10)
  - 1.1.1 Catalog (1)
  - 1.1.2 Entry (1)
- 1.2 Title (1)
- 1.3 Language (10)
- 1.4 Descriptor (10)
- 1.5 Keyword (10)
- 1.6 Coverage (10)

2 Life Cycle (1)
- 2.3 Contribute (30)
  - 2.3.1 Role (1)
  - 2.3.2 Entity (40)
  - 2.3.3 Date (1)

3 Meta-Metadata (1)
- 3.1 Identifier (10)
  - 3.1.1 Catalogue (1)
  - 3.1.2 Entry (1)
- 3.2 Contribute (10)
  - 3.2.1 Role (1)
  - 3.2.2 Entity (10)
  - 3.2.3 Date (1)
- 3.3 Metadata Schema (10)
- 3.4 Language (1)

4 Technical (1)
- 4.1 Format (40)
- 4.2 Size (1)
- 4.3 Location (10)
- 4.7 Duration (1)

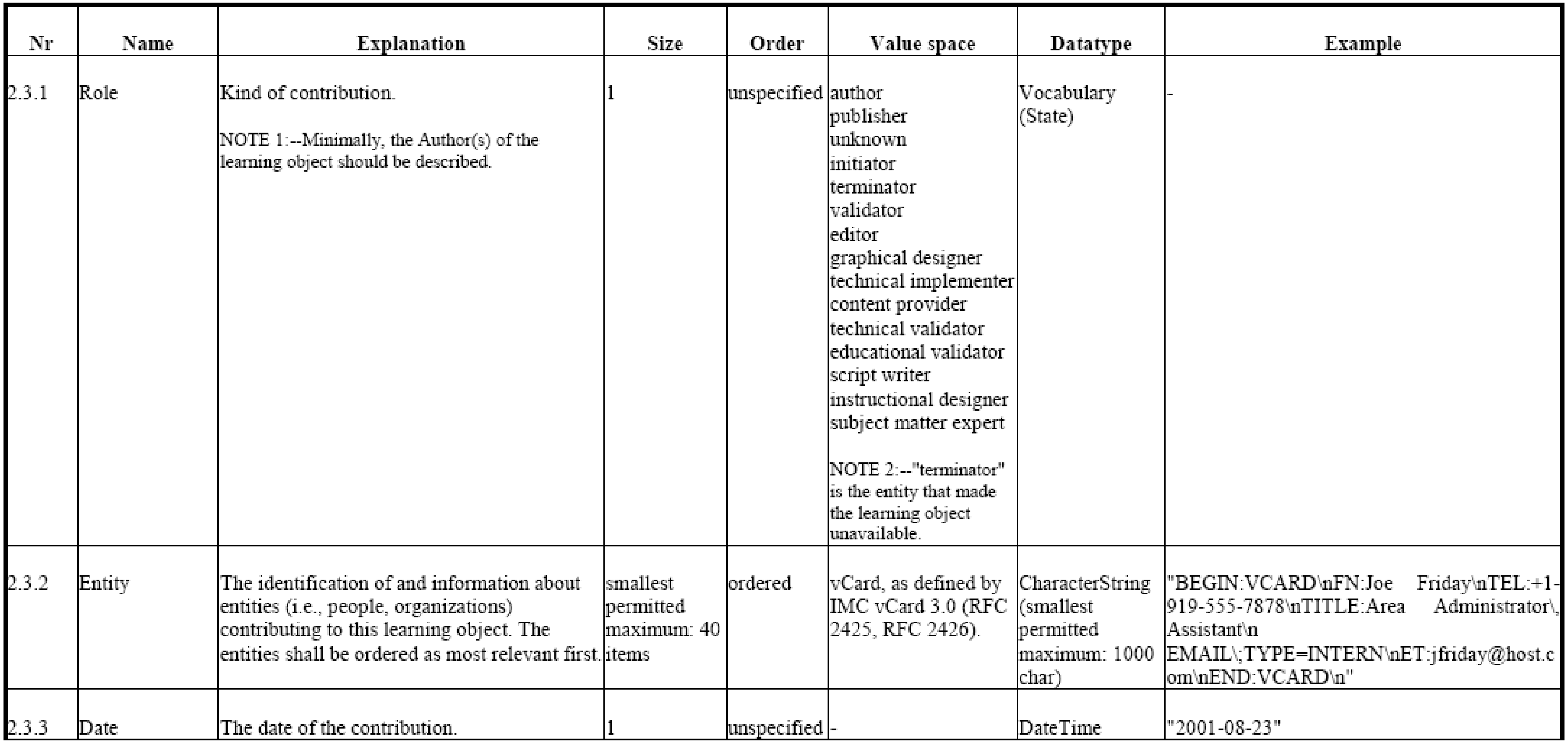
5 Educational (100)
- 5.2 Learning Resource Type (10)
- 5.8 Difficulty (1)
- 5.9 Typical Learning Time (1)
- 5.10 Description (10)
- 5.11 Language (10)
- *5.12 Language Specifics* (1)
  - 5.12.1 Accent (5)
  - 5.12.2 Region (5)
  - 5.12.3 Register (1)
- 5.13 Suggestions for further use (1)
- 5.14 Information for Users (1)
- *5.15 Description For Learner* (1)
- *5.16 Genre* (10)
- *5.17 Language Function* (5)
- *5.18 Language Mode* (5)
- *5.19 Subtitles/Transcript Included* (1)
  - 5.19.1 Subtitles (1)
  - 5.19.2 Language (1)
- *5.20 Document Length (1)*
- 5.21 Instruction Language (1)

6 Rights (1)
- 6.2 Copyright and Other restrictions (1)
- 6.3 Description (1)

7 Relation (100)
- 7.1 Kind (1)
- 7.2 Resource (1)
  - 7.2.1 Identifier (10)
    - 7.2.1.1 Catalogue (1)
    - 7.2.1.2 Entry (1)
  - 7.2.2 Description (10)

9 Classification (40)
- 9.1 Purpose (1)
- 9.3 Description (1)

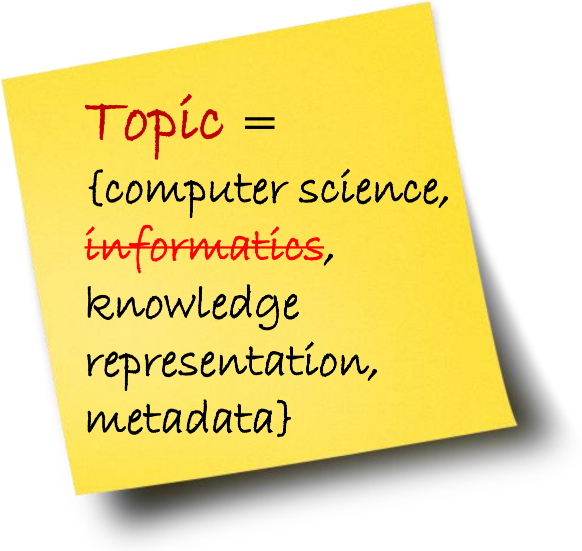
Example: IEEE LOM



Nr	Name	Explanation	Size	Order	Value space	Datatype	Example
2.3.1	Role	Kind of contribution. NOTE 1:--Minimally, the Author(s) of the learning object should be described.	1	unspecified	author publisher unknown initiator terminator validator editor graphical designer technical implementer content provider technical validator educational validator script writer instructional designer subject matter expert NOTE 2:--"terminator" is the entity that made the learning object unavailable.	Vocabulary (State)	-
2.3.2	Entity	The identification of and information about entities (i.e., people, organizations) contributing to this learning object. The entities shall be ordered as most relevant first.	smallest permitted maximum: 40 items	ordered	vCard, as defined by IMC vCard 3.0 (RFC 2425, RFC 2426).	CharacterString (smallest permitted maximum: 1000 char)	"BEGIN:VCARD\nFN:Joe Friday\nTEL:+1-919-555-7878\nTITLE:Area Administrator\, Assistant\n EMAIL\;TYPE=INTERN\nET:jfriday@host.com\nEND:VCARD\n"
2.3.3	Date	The date of the contribution.	1	unspecified	-	DateTime	"2001-08-23"

… + controlled vocabularies

- A closed list of named subjects, which can be used for classification
- Metadata field values are restricted to a list of terms (selected by experts)

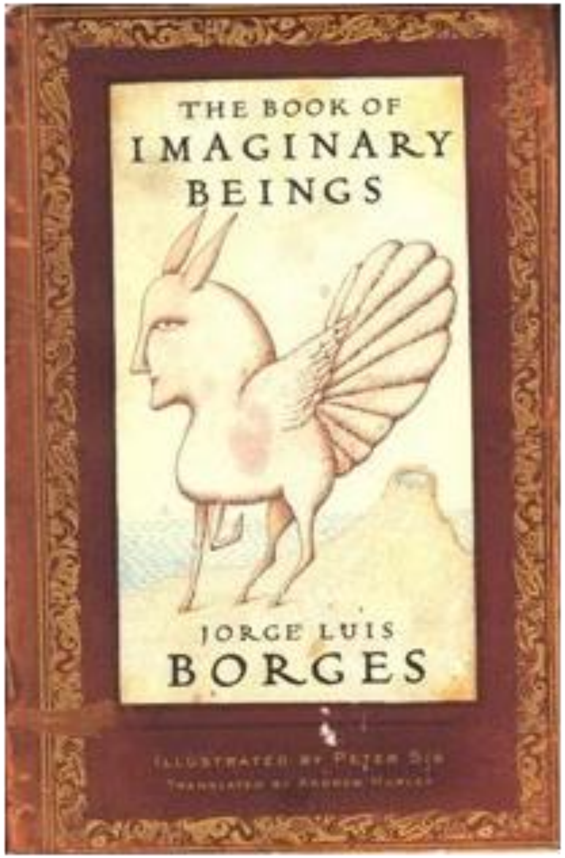


Topic =
{computer science,
~~informatics~~,
knowledge
representation,
metadata}

Subject-based classification

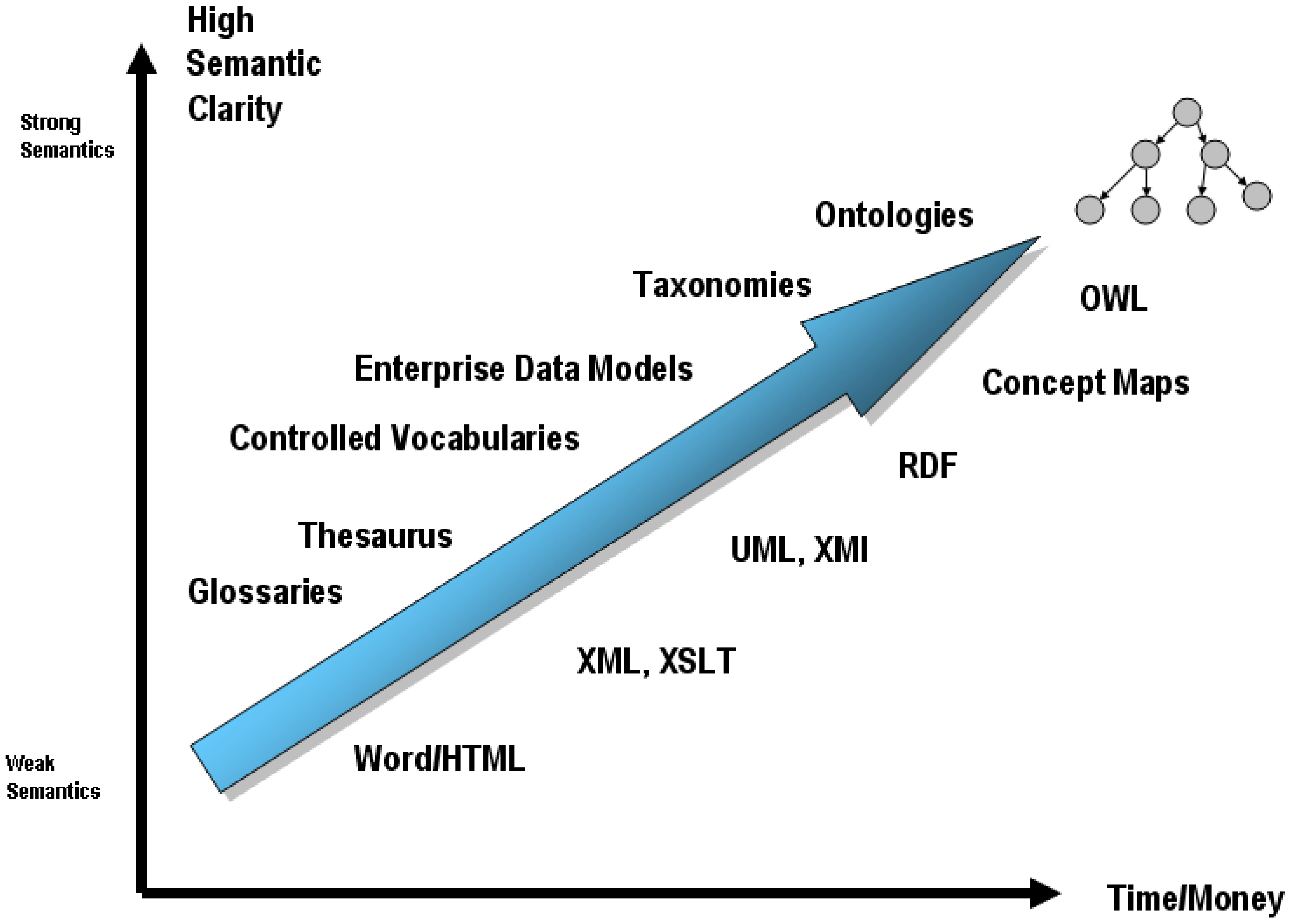
- Any form of content classification that groups objects by their subjects
  - e.g the use of keywords to classify papers
- Metadata fields describe what the objects are about by listing discrete subjects inside a subject-based classification
- Important: difference between describing the objects being classified and describing the subjects used to classify them
  - Metadata describe objects
  - Subject-based classification is the approach to describe subject

Subject-based classification



those that belong to the Emperor,
embalmed ones,
those that are trained,
suckling pigs,
mermaids,
fabulous ones,
stray dogs,
those included in the present classification,
those that tremble as if they were mad,
innumerable ones,
those drawn with a very fine camelhair brush,
others,
those that have just broken a flower vase,
those that from a long way off look like flies.

Subject-based classification

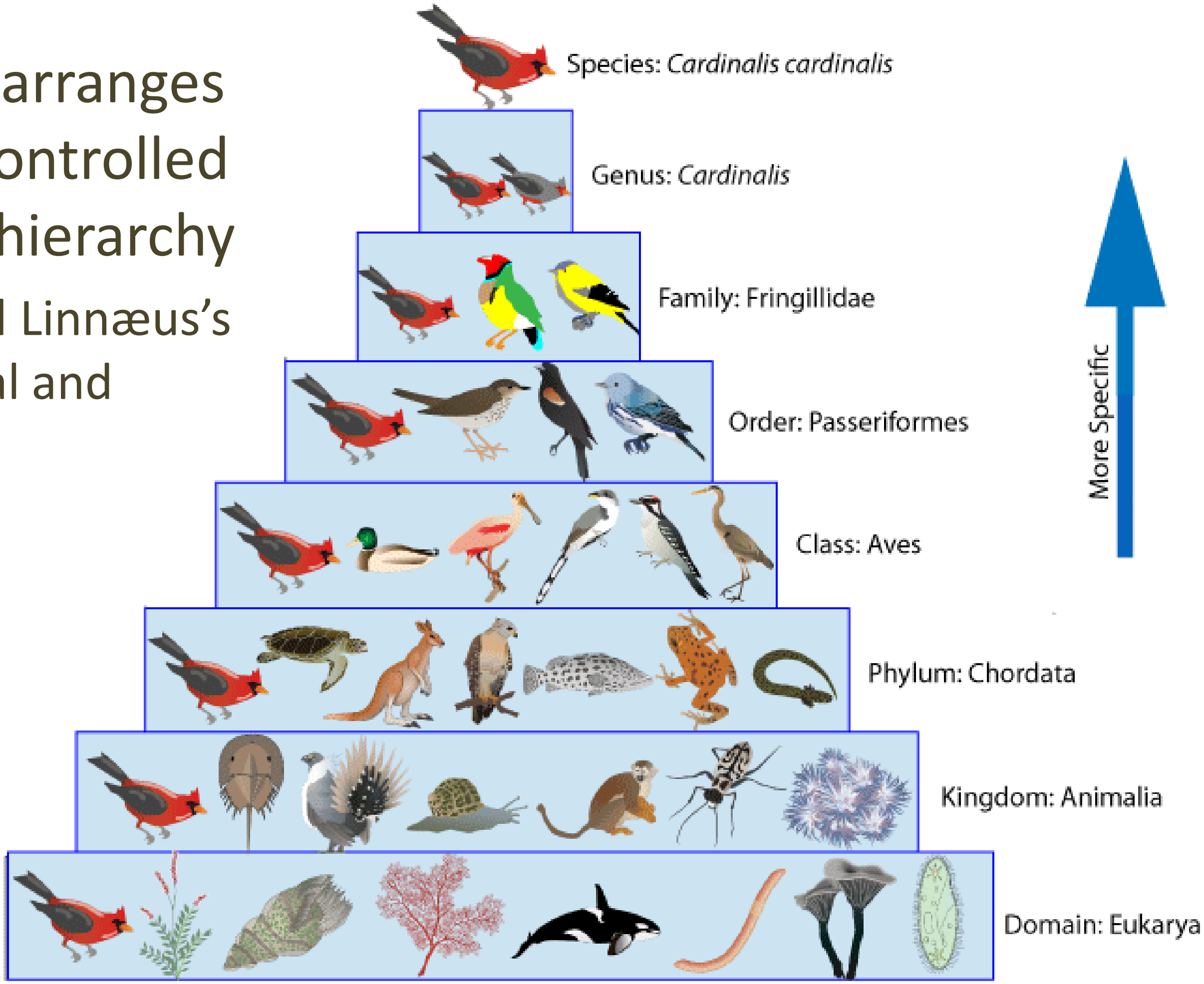


Controlled vocabulary

- Goal
  - Prevent authors from defining terms that are meaningless, too broad or too narrow
  - Prevent authors from misspelling
  - Prevent different authors from choosing slightly different forms of the same term
- Simplest form: list of terms (or “pick list”)
- Reduces ambiguity inherent in normal human languages
- Solves the problems of homographs, homonyms, synonyms and polysemes by ensuring
  - That each concept is described using only one authorized term
  - That each authorized term in the controlled vocabulary describes only one concept

Taxonomy

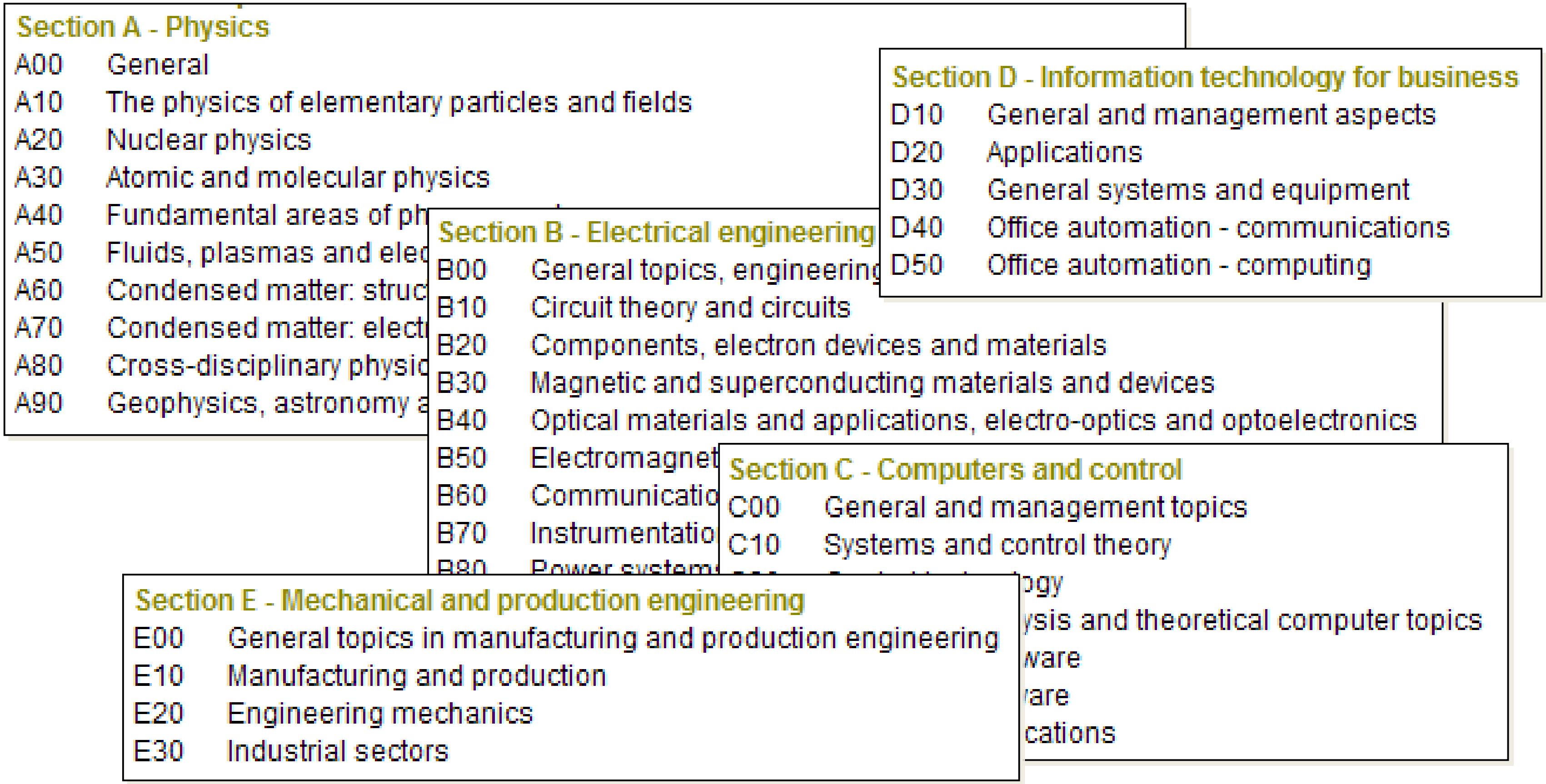
- Subject-based classification that arranges the terms in the controlled vocabulary into a hierarchy
  - Dates back to Carl Linnæus's work on zoological and botanical classification (18th century)



Taxonomy example: INSPEC

- Objective: to index quality research literature in physics and engineering

http://www.theiet.org/publishing/inspec/index.cfm



Section A - Physics

A00	General
A10	The physics of elementary particles and fields
A20	Nuclear physics
A30	Atomic and molecular physics
A40	Fundamental areas of ph
A50	Fluids, plasmas and elec
A60	Condensed matter: struc
A70	Condensed matter: electr
A80	Cross-disciplinary physic
A90	Geophysics, astronomy a

Section B - Electrical engineering

B00	General topics, engineering
B10	Circuit theory and circuits
B20	Components, electron devices and materials
B30	Magnetic and superconducting materials and devices
B40	Optical materials and applications, electro-optics and optoelectronics
B50	Electromagnet
B60	Communicatio
B70	Instrumentatio
B80	Power system

Section D - Information technology for business

D10	General and management aspects
D20	Applications
D30	General systems and equipment
D40	Office automation - communications
D50	Office automation - computing

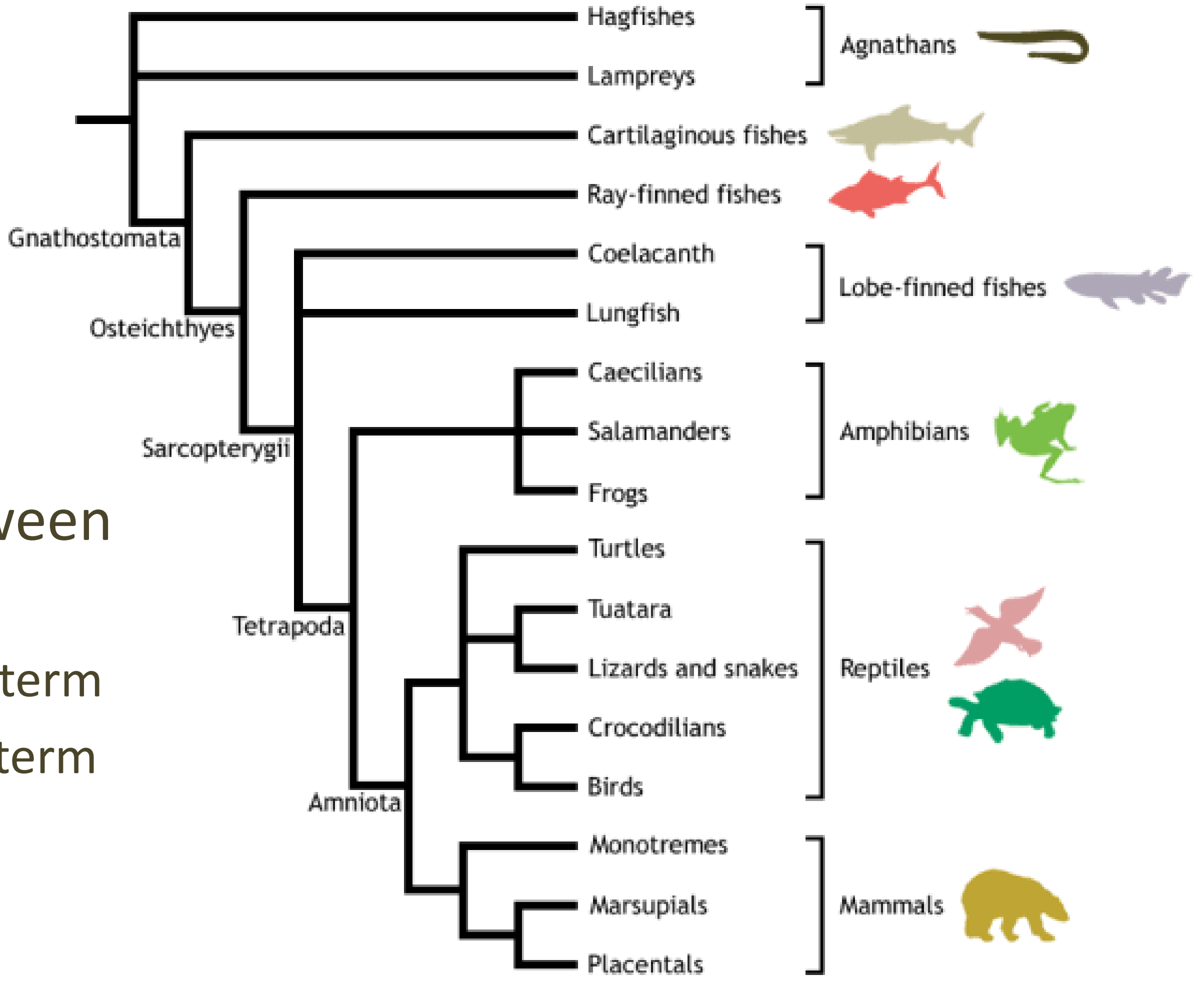
Section C - Computers and control

C00	General and management topics
C10	Systems and control theory
	ogy
	ysis and theoretical computer topics
	ware
	vare
	cations

Section E - Mechanical and production engineering

E00	General topics in manufacturing and production engineering
E10	Manufacturing and production
E20	Engineering mechanics
E30	Industrial sectors

Limit of taxonomies

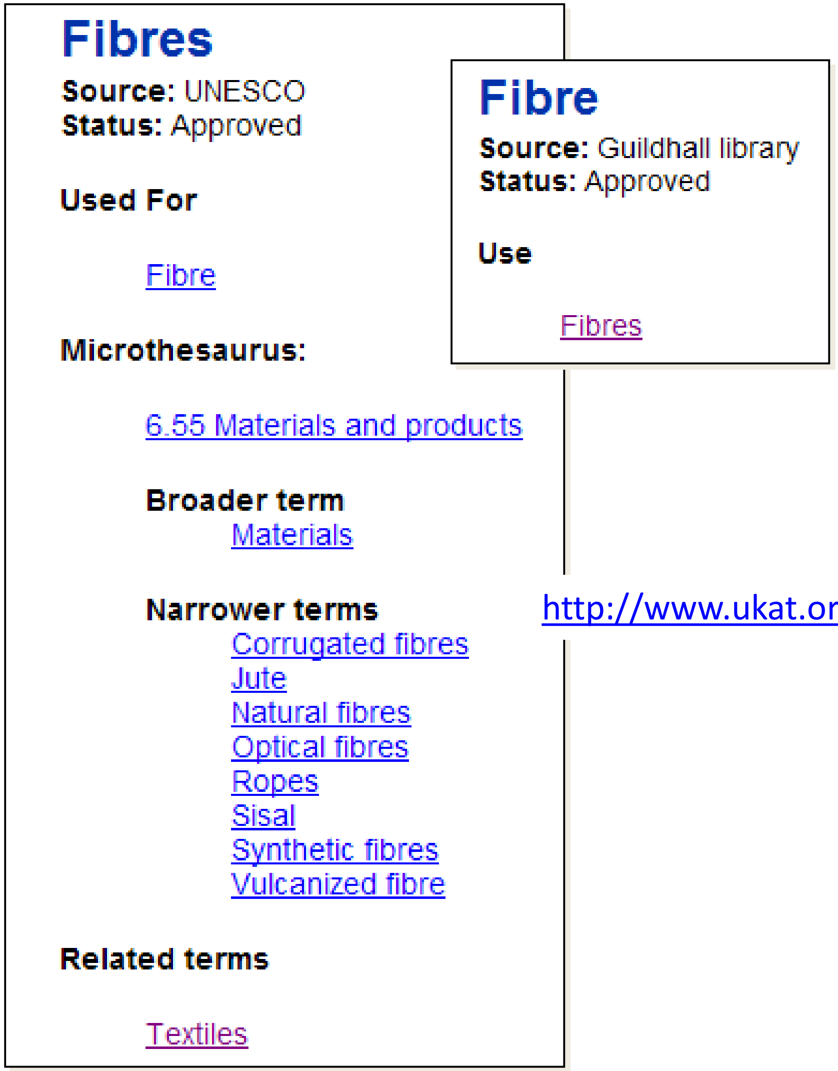


- Only two kinds of relationships between terms
  - Parent = broader term
  - Child = narrower term

Thesaurus

- Extends taxonomies
  - subjects are arranged in a hierarchy
- Other statements can be made about the subjects
  - BT – broader term
  - NT - narrower term (inverse of BT)
  - SN – scope note
  - USE
  - UF – used for (inverse of USE)
  - TT – top term
  - RT – related term

Thesaurus example



Fibres

Source: UNESCO
Status: Approved

Used For

Fibre

Microthesaurus:

6.55 Materials and products

Broader term

Materials

Narrower terms

- Corrugated fibres
- Jute
- Natural fibres
- Optical fibres
- Ropes
- Sisal
- Synthetic fibres
- Vulcanized fibre

Related terms

Textiles

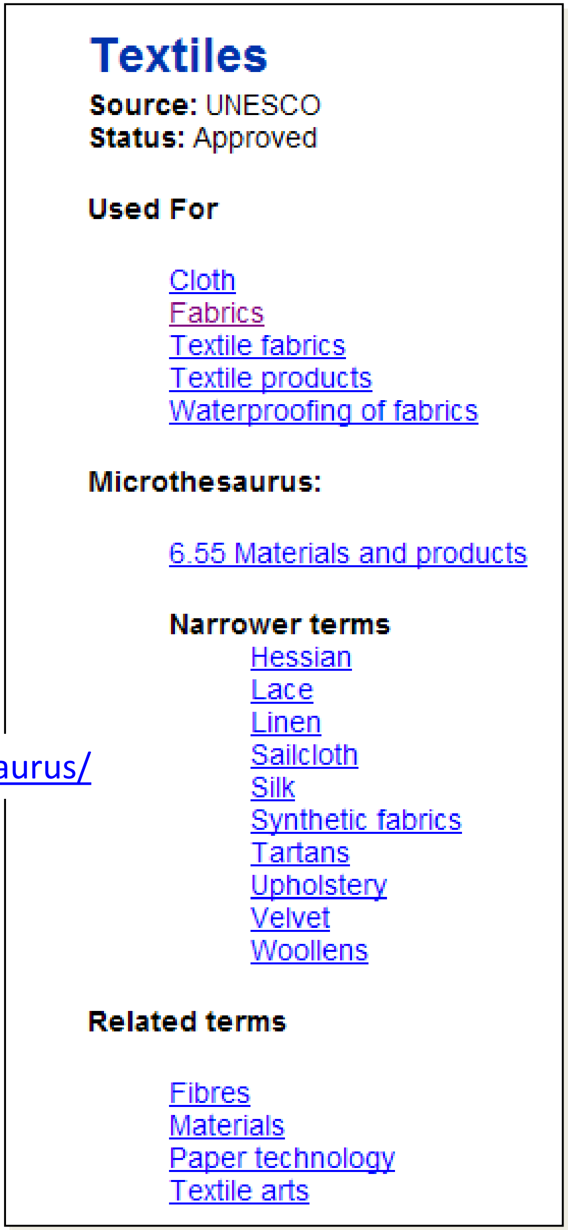
Fibre

Source: Guildhall library
Status: Approved

Use

Fibres

http://www.ukat.org.uk/thesaurus/



Textiles

Source: UNESCO
Status: Approved

Used For

- Cloth
- Fabrics
- Textile fabrics
- Textile products
- Waterproofing of fabrics

Microthesaurus:

6.55 Materials and products

Narrower terms

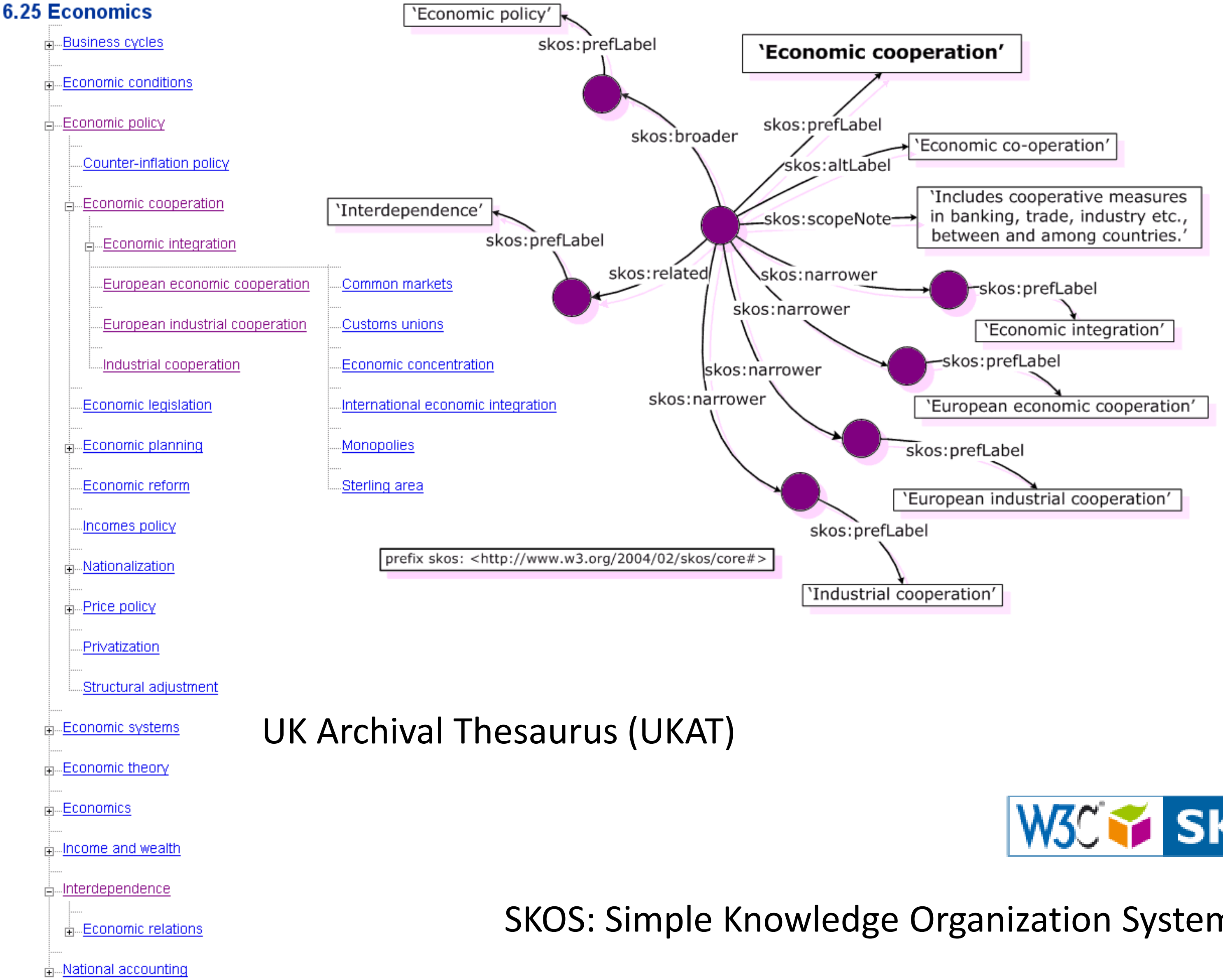
- Hessian
- Lace
- Linen
- Sailcloth
- Silk
- Synthetic fabrics
- Tartans
- Upholstery
- Velvet
- Woollens

Related terms

- Fibres
- Materials
- Paper technology
- Textile arts

UKAT

RDF / SKOS



6.25 Economics

- Business cycles
- Economic conditions
- Economic policy
  - Counter-inflation policy
  - Economic cooperation
    - Economic integration
      - Common markets
      - Customs unions
      - Economic concentration
      - International economic integration
      - Monopolies
      - Sterling area
    - European economic cooperation
    - European industrial cooperation
    - Industrial cooperation
  - Economic legislation
  - Economic planning
  - Economic reform
  - Incomes policy
  - Nationalization
  - Price policy
  - Privatization
  - Structural adjustment
- Economic systems
- Economic theory
- Economics
- Income and wealth
- Interdependence
  - Economic relations
- National accounting

prefix skos: <http://www.w3.org/2004/02/skos/core#>

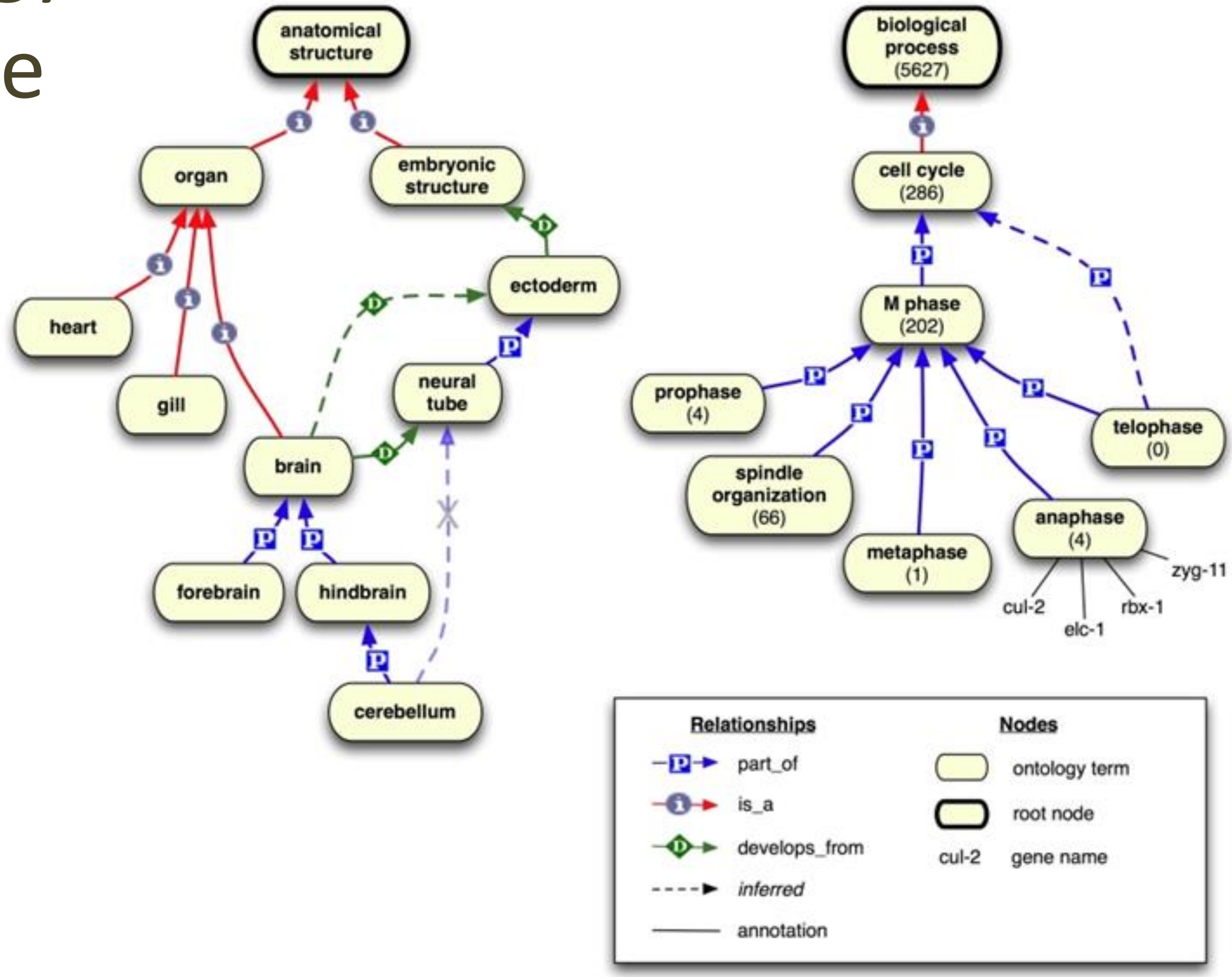
UK Archival Thesaurus (UKAT)

SKOS: Simple Knowledge Organization System

Ontology

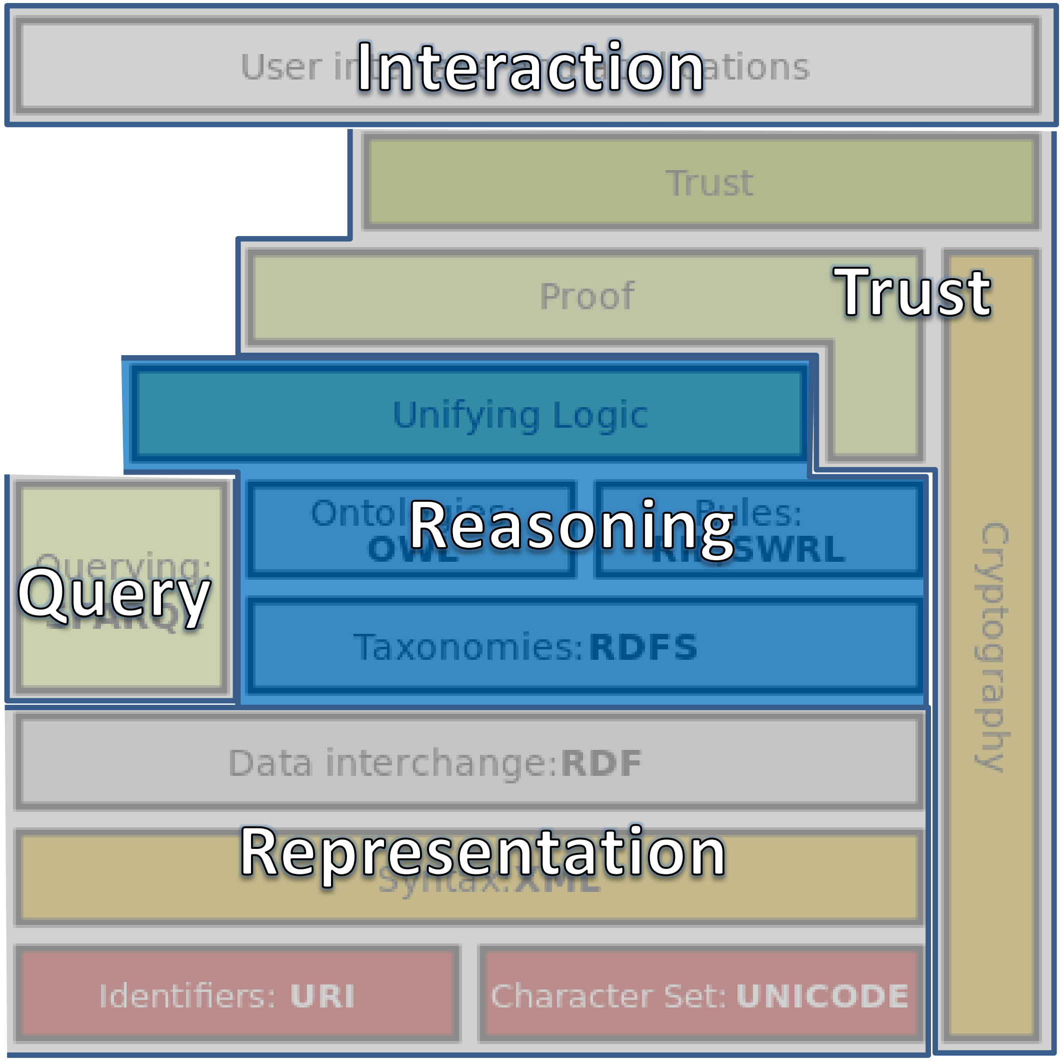
- Model for describing the world that consists of a set of types, properties, and relationships
- Extends the other subject-based classification approaches
  - Has open vocabularies
  - Has open relationship types (not just BT/NT, RT and USE/UF)

Ontology example



Semantically rich descriptions to support search

- Step 2: reasoning
- Ontologies



References

- W3C Semantic Web
  - https://www.w3.org/standards/semanticweb/
- W3C Tutorial on Semantic Web
  - https://www.w3.org/Consortium/Offices/Presentations/RDFTutorial/
- Lee Feigenbaum, “The Semantic Web Landscape”
  - http://www.slideshare.net/LeeFeigenbaum/cshals-2010-w3c-semanic-web-tutorial

License