3. Knowledge Representation, Ontology & Metadata

- Next Tutorial: 21.11
- Team up for the extended abstract!
  - Deadline: 21.11

3. Knowledge representation, Ontology & Metadata

3.1 Introduction

3.2 Classifications & Ontologies

3.3 Indexing: Words and thesauri

3.4 Metadata & Semantic Web

3.5 Vector models

3. Knowledge representation, Ontology & Metadata

- What is Knowledge Representation?
  - Surrogate
  - Ontological commitments
  - Intelligent reasoning
  - Medium for efficient computation
  - Medium for human expression

Davis et al. 1993 (MIT AI lab)

(http://groups.csail.mit.edu/medg/ftp/psz/k-rep.html)

3.1 Introduction

What are digital libraries for?
- providing **controlled access** to information

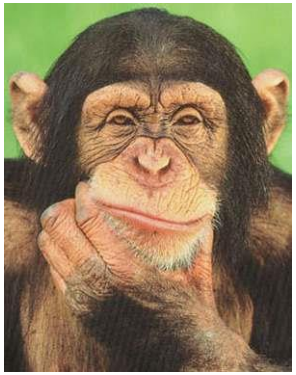
3.1 Introduction

Common problem with DLs: queries often returns either nothing or thousands of answers.

Challenge: How to organize and represent knowledge to support effective search?

3.1 Introduction

Thoughts:



A single knowledge representation schema that let us put each idea in one place, and if the users knew this schema …. It would be perfect!!!

But…

in practice, it seems unlikely that any single knowledge representation schema will serve all purposes… the more detailed it is, the less likely it is that two different people will come up with the same place for the same document, and the less detailed it is, the less resolving power it has and the less useful it is…

3.1 Introduction

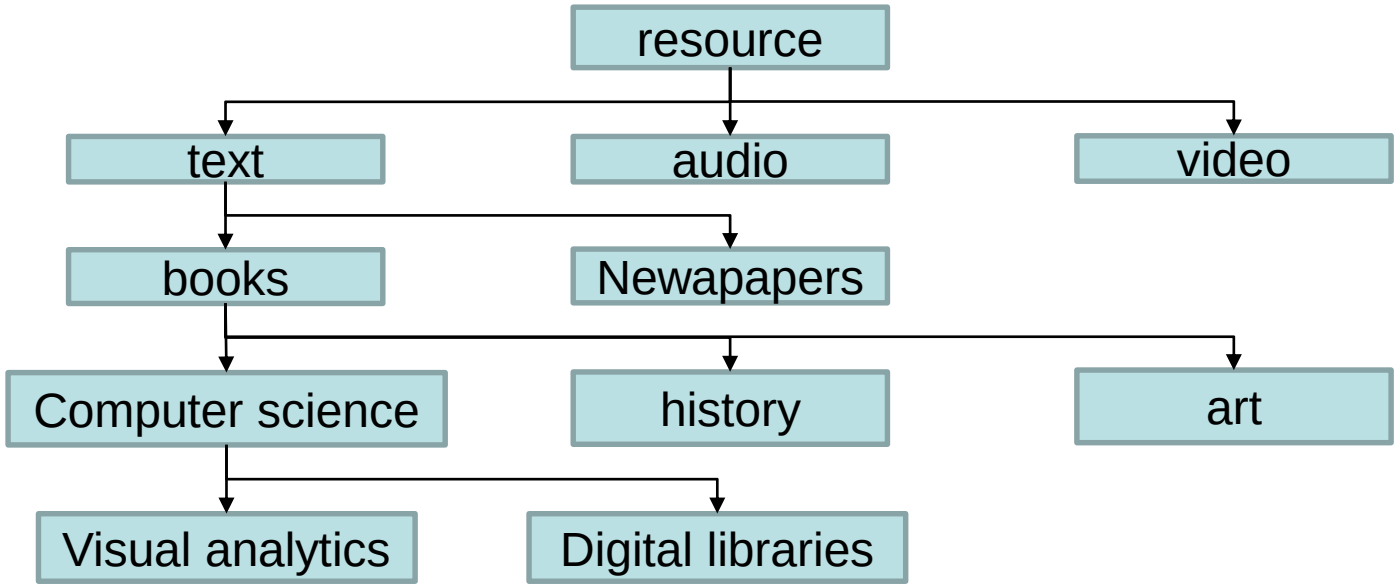
Knowledge representation schemas to be discussed in the following:

- Library Classifications for knowledge labeling
- Metadata to aid the identification, description and location of resources
- Thesauri to disambiguate word meaning
- Ontologies to specify relevant concepts and the semantics relationships
- Semantic Web to represent knowledge in patterns of interconnected nodes
- Algebraic (vector) model for representing text documents

3. Knowledge representation, Ontology & Metadata

3.2 Classification & Ontology

- A **classification** is a structure that organizes concepts into a hierarchy, possibly in a scheme of facets. → **bibliographic system**
- Each entity (book, media …) must be put in only one category "**mark and park**"



3.2 Classification & Ontology

- **Ontology** is a specification of a conceptualization (Gruber, 1993).

  Major components:

  Concepts, i.e. human, animal, food, table, movie, etc.

  Instances, i.e. Tom is an instance of concept “person”.

  Properties, i.e. a human has properties of gender, height, weight, father, mother, etc.

  Relations, i.e. University of Konstanz is located in Konstanz.

  Rules. If someone is married, then he/she should have a spouse.

3.2 Classification & Ontology

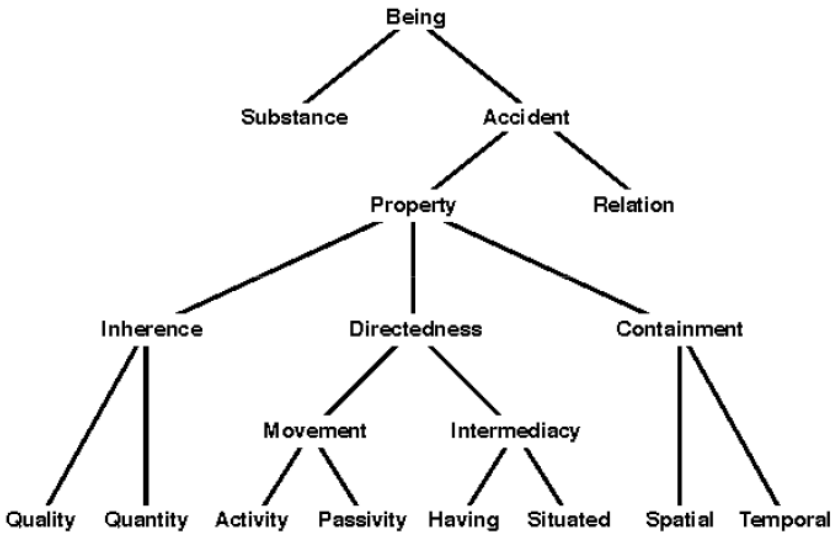
Trace back to Aristotle (384 BC – 322 BC) - he listed ten categories that all things of the world should belong to (in his work “categories”), and illustrated the ontology (in “Science of Being”)

Aristotle's 10 categories

1. substance,
2. quantity,
3. quality,
4. relation,
5. place,
6. time,
7. position,
8. state,
9. action, and
10. passion.

For example, "A five-foot tall (quantity) man (substance) who was a thinker (quality) sat (position) on a bus (place) one morning (time), feeling hungry (state), but continuing to do a crossword puzzle (action) enthusiastically (passion)."

Aristotle's ontology



3.2.1 Library Classification Systems

Early classification systems

The library of Congress classification system

- Before 1812 (Francis Bacon)
- After 1814 (Thomas Jefferson)

To 1812 (Based on Bacon)		*From 1814 on (Jefferson)*			
1.	Sacred history	1.	History, ancient	23.	Foreign laws
2.	Ecclesiastical history	2.	Modern history except UK, US	24.	Politics
3.	Civil history	3.	Modern history, British Isles	25.	Arithmetic
4.	Geography, travels	4.	Modern history, America	26.	Geometry
5.	Law	5.	History, ecclesiastical	27.	Mathematical physics: mechanics, optics
6.	Ethics	6.	Physics, natural philosophy	28.	Astronomy
7.	Logic, rhetoric, criticism	7.	Agriculture	29.	Geography
8.	Dictionaries, grammars	8.	Chemistry	30.	Fine arts, architecture
9.	Politics	9.	Surgery	31.	Gardening, painting, sculpture
10.	Trade & commerce	10.	Medicine	32.	Music
11.	Military & naval tactics	11.	Anatomy	33.	Poetry, epic
12.	Agriculture	12.	Zoology	34.	Romance, fables
13.	Natural history	13.	Botany	35.	Pastorals, odes
14.	Medicine, surgery, chemistry	14.	Mineralogy	36.	Didactic
15.	Poetry and drama, fiction	15.	Technical arts	37.	Tragedy
16.	Arts & sciences & miscellaneous	16.	Ethics	38.	Comedy
17.	Gazettes (newspapers)	17.	Religion	39.	Dialogue (epistolary)
18.	Maps	18.	Equity (law)	40.	Logic, rhetoric
		19.	Common law	41.	Criticism, theory
		20.	Commercial law	42.	Criticism, bibliography
		21.	Maritime law	43.	Criticism, languages
		22.	Ecclesiastical law	44.	Polygraphical

The British Museum Library classification system

- Before 1808 (14 headings, some reflect the source rather than the content of books)
- After 1808: Antonio Panizzi's system

  The system is extremely idiosyncratic. e.g., "evidence for and against Christianity", "marriage – female suffering", "morality of war, cruelty to animals, dueling"

1. Philology, Memoirs of Academies, Classics
2. Cracherode Library
3. Poetry, Novels, Letters, Polygraphy
4. History (ancient), Geography, Travels
5. Modern History
6. Modern History, Biography, Diplomacy, Heraldry, Archaeology, Numismatics, Bibliography
7. Medicine, Surgery, Trade & Commerce, Arts, Mathematics, Astronomy
8. Medicine, Natural History
9. Politics, Philosophy, Chemistry, Natural History
10. Ecclesiastical History, Jurisprudence, Divinity
11. Divinity
12. Sermons, Political tracts, King's pamphlets
13. Acta Sanctorum, Musgrave Biographical Collection, Music
14. Parliamentary records, Gazettes, Newspapers

3.2.1 Library Classification Systems

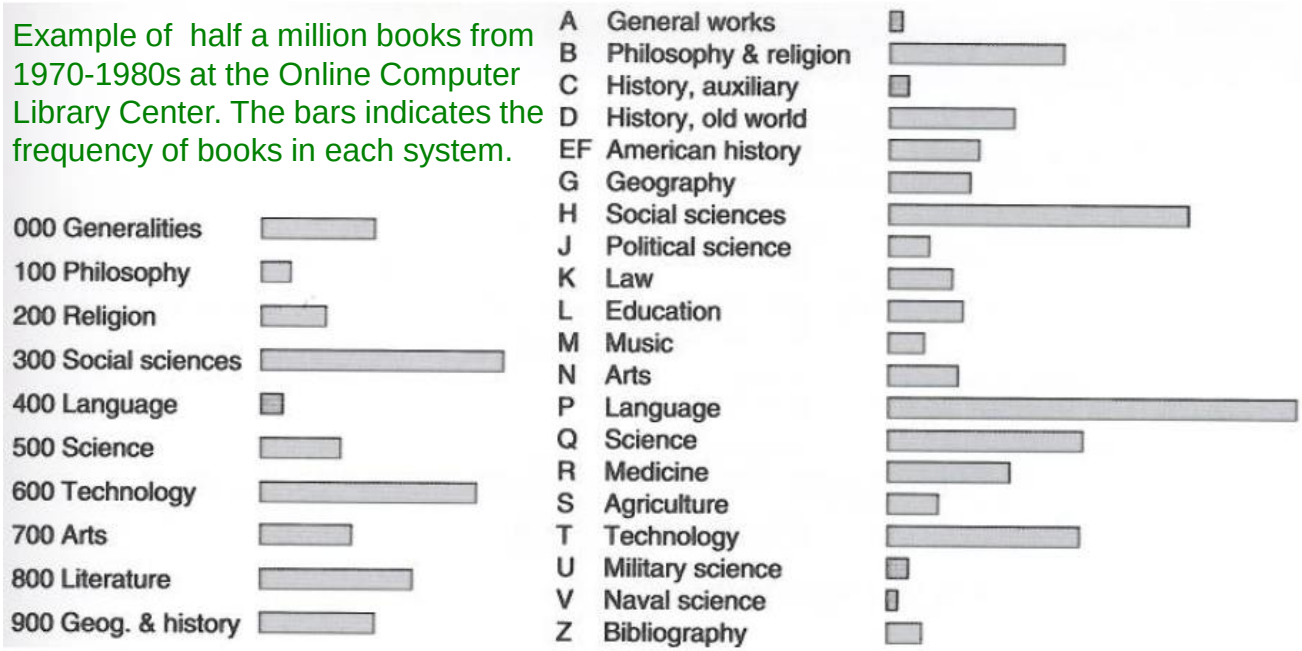
End of 19th century – major classification systems started

USA

- **Dewey system** 1876 (Melvil Dewey) - preferred by public /smaller libraries
- **New Library of Congress System** 1898-1920(Charles Cutter) - preferred by research libraries /larger libraries

Dewey System | New Library of Congress System

Example of half a million books from 1970-1980s at the Online Computer Library Center. The bars indicates the frequency of books in each system.



Europe

- **Universal Decimal Classification** - Resembles Dewey in many ways but is maintained separately

3.2.1 Library Classification Systems

Different classifiers may make different decisions on the same book

"Scalded to death by the steam "

Published in both US and UK, book of songs about railway accidents.

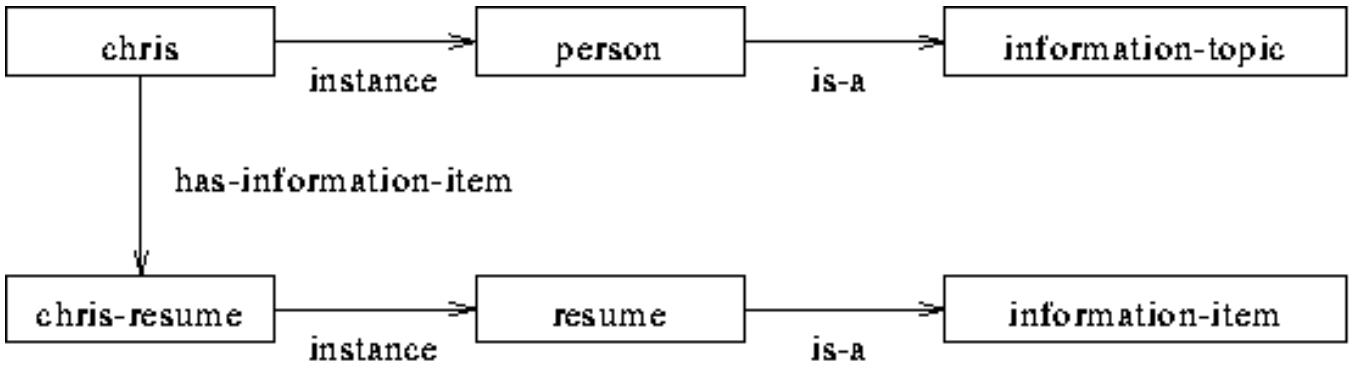
	The US system	The UK system
Primary subject	Songs	Railways
Sub category	Railways	Music

3.2.1 Classification System for Websites

- Collectively edited web directories (DMOZ)
- Google Directory was abandoned in 2011

3.2.2 Digital Library Ontologies

- Formal Ontologies for DLs
  - ☑ specify relevant concepts – the types of things and their properties – and the semantics relationships that exist between those concepts in a particular domain.
  - ☑ mathematically well-defined syntax and semantics to describe such concepts, properties, and relationships precisely

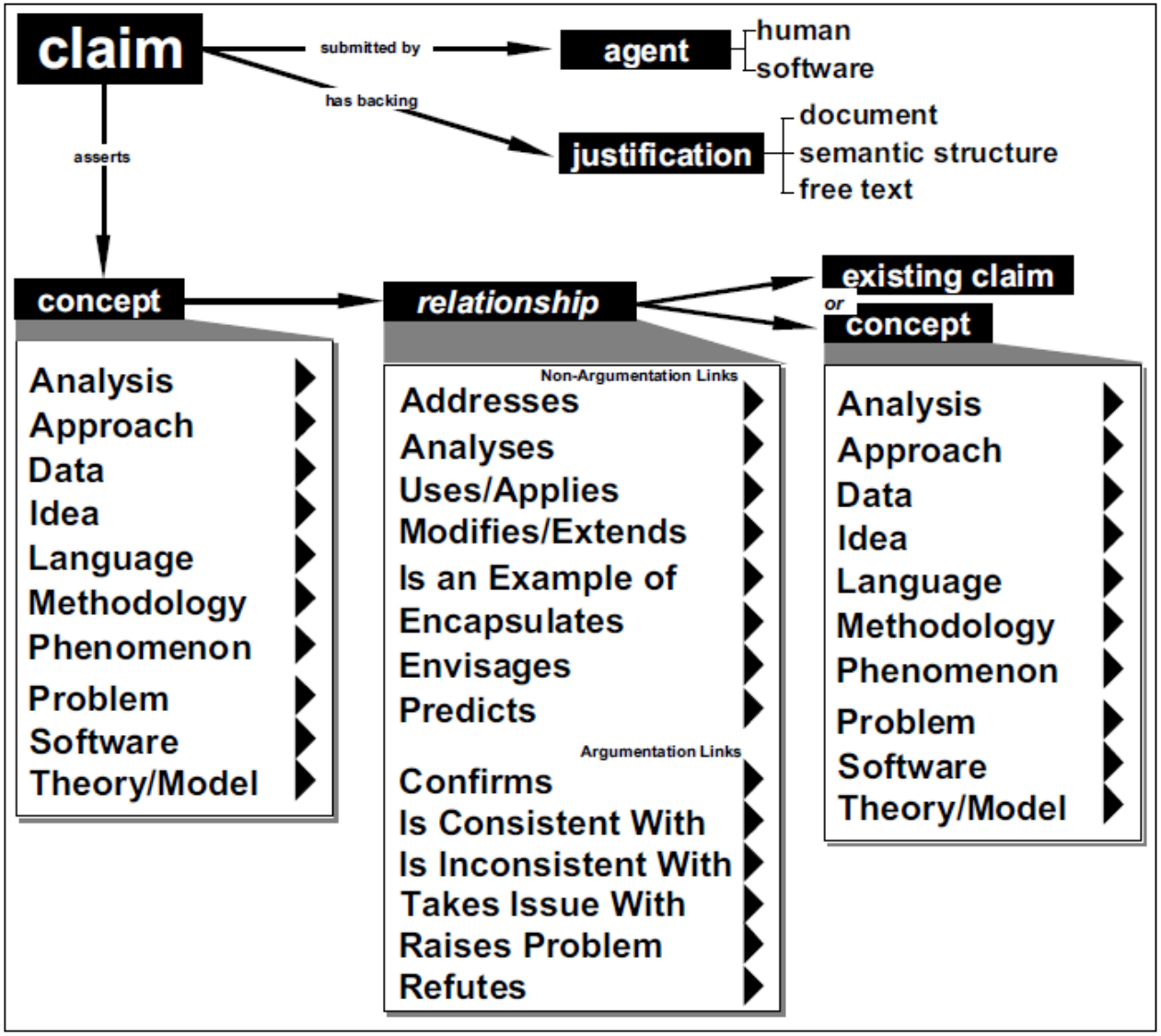


3.2.2 Library Ontologies

Structure of an ontology - typically an ontology has two distinct components:

- Names for important concepts in the domain
  - *Elephant* is a concept whose members are a kind of animal
  - *Herbivore* is a concept whose members are exactly those animals who eat only plants or parts of plants
  - *Adult_Elephant* is a concept whose members are exactly those elephants whose age is greater than 20 years
- Background knowledge/constraints on the domain
  - *Adult_Elephant*s weigh at least 2,000 kg
  - All *Elephant*s are either *African_Elephant*s or *Asian_Elephant*s
  - No individual can be both a *Herbivore* and a *Carnivore*

3.2.2 Digital Library Ontologies



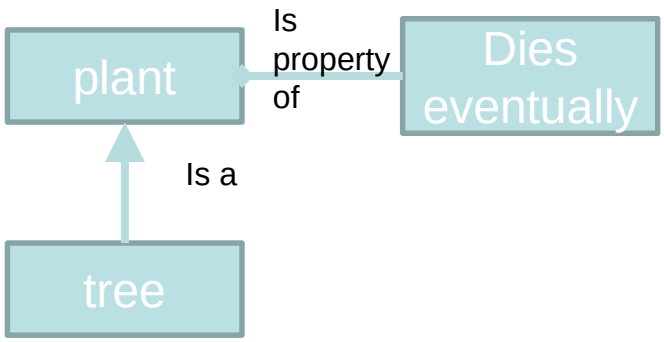
3.2.2 Digital Library Ontologies

- MOPs (memory organization processes), Rodger Schank & students- 1980's
  - early attempt to develop knowledge base
  - represent certain standard scenarios for common situations
  - e.g., "restaurant script", "natural disaster"

- CYC, Doug Lenat, 1984- 2004
  - attempt to produce a comprehensive ontology and knowledgebase of everyday common sense knowledge
  - idea: write down all common sense knowledge automatic reasoning

    Rule 1. "Every tree is a plant", Rule 2. "Plants die e

    Reasoning: All trees die eventually.



  - 100,000 concepts and 1 million rules.
  - no unifying overall ontology, broken down into a number of "micro-theories"

3. Knowledge representation, Ontology & Metadata

4.1 Introduction

4.2 Classifications & Ontologies

4.3 Indexing: words and thesauri

4.4 Metadata & Semantic Web

4.5 Vector models

3.3 Indexing: words and thesauri

- A **dictionary** is a listing of words and phrases giving information such as spelling, morphology and part of speech, senses, definitions, usage, origin, and equivalents in other languages (bi- or multilingual dictionary).

- A **thesaurus** is a **structure** that manages the complexities of terminology and provides conceptual relationships, ideally through an embedded classification/ontology.

- A thesaurus may specify descriptors authorized for indexing and searching. These descriptors form a **controlled vocabulary** (**authority list**, **index language**).

3.3.1 Words

Techniques to find out the exact meaning of words in a text

- **Geoffrey Sampson** – statistical method based on the probabilities of certain word sequences

Time flies like an arrow.

("flies" can't be a noun as it is behind a noun)

- **Garside 1987** – pair of words at a time
- **Ken Church 1988** – trigrams, better results
- **Lesk 1986** – count overlaps between the definitions of different sense of nearby word ***pine cone***

word	sense	count	Sense definition
pine	1*	7	Kinds of evergreen tree with needle -shaped evergreen(1) tree(6)
	2	1	Pine pine(1)
	3	0	Waste away through sorrow or illness
	4	0	/pine for something; pint to do something

word	sense	count	Sense definition
cone	1	0	Solid body which narrows to a point from a
	2	0	sth of this shape whether solid or hollow
	3*	8	Fruit of certain evergreen trees (fir, pine, evergreen(1) tree(6) pine(1)

- Stop words, weighting of words…

3.3.2 Thesauri

Statistical methods don't always work…

Thesauri:

- Vocabulary confusion
- User orientation in a concept space

Alcohol and Other Drug Thesaurus (AOD Thesaurus)
(US Nat. Inst. of Alcohol Abuse and Alcoholism)
http://etoh.niaaa.nih.gov/AODVol1/Aodthome.htm

Medical Subject Headings (MeSH) and Unified Medical Language System (UMLS)
(US National Library of Medicine)
www.nlm.nih.gov/mesh/meshhome.html, www.nlm.nih.gov/mesh/MBrowser.html, www.nlm.nih.gov/research/umls/umlsmain.html, http://umlsinfo.nlm.nih.gov

Art and Architecture Thesaurus (AAT)
(Getty Foundation)
http://www.getty.edu/research/tools/vocabulary/aat/index.html

Still 50% of the time experienced indexers disagree on which terms to assign, 85% of novice indexers disagree…

3.3.2 Thesauri – example queries

Queries: Synonym expansion and Hierarchic expansion

Query 1.	Drug use by teenagers
Query 1.1	teenage* AND drug*
Query 1.2	Synonym expansion for teenage* (teenage* OR teen OR teens OR youth* OR adolescent* OR kid* OR "high school") AND drug*
Query 1.3	In addition, synonym expansion and hierarchic expansion for drug* (teenage* OR teen OR teens OR youth* OR adolescent* OR kid* OR "high school") AND (drug* OR substance* OR alcohol OR nicotine OR smoking OR cigarette* OR mari*una OR cocaine OR crack OR heroin)
Query 1.4	Query more narrowly focused (teenage* OR teen OR teens OR youth* OR adolescent* OR kid* OR "high school") AND (cocaine OR crack OR heroin)

3.3.2 Thesauri – query results

Query 1.1. teenage* AND drug* (AltaVista)

About 30 documents match your query.

1. CEIDA Druglinks - Info Centre - PARENTS TALKING TO TEENAGERS ABOUT DRUGS
What do parents want from their teenagers? Basically, parents want: To know your kids are alright and not in danger. To know your kids think you're OK...
http://www. ceida. net. au/info_centre/drug~myths/what_do. html - size 3K - 21-May-97 - English

2. CEIDA Druglinks - Info Centre - PARENTS TALKING TO TEENAGERS ABOUT DRUGS
Better Ways of Communicating. Different points of view Communication is the key to resolving problems, if they exist. Or to finding out if they exist....
http1A~www. ceida. net. au/info_centre/drug~myths/better.html - size 9K - 21-May-97 - English

3. Testimony of Donna E. Shalala, Secretary of HHS on Teenage Drug Use
Testimony of Donna E. Shalala, Secretary of Health and Human Services on Teenage Drug Use. Testimony of. Donna E. Shalala. Secretary of Health and Human...
http://www.apa.org/ppo/shalala.html - size 15K - 13-Sep-96 - English

4. Statement of Senator Richard C. Shelby on Teenage Drug Use
Statement of Senator Richard C. Shelby on Teenage Drug Use. Statement of. U.S. Senator Richard C. Shelby. Before The. Senate Judiciary Committee Hearing..
http://www.apa.org/ppo/shelbyhtml - size 3K - 13-Sep-96 - English

5. Testimony of John P. Walters on Teenage Drug Use
Testimony of John P. Walters, President of The New Citizenship Project, on Teenage Drug Use. Testimony by. John P. Walters* President of the New...
http://www.apa.org/ppo/walters.html - size 28K - 13-Sep-96 - English

6. Drug Use Rises for Teenagers
Parent News for November 1996. Of Interest. Drug Use Rises for Teenagers. by Anne S. Robertson. A recent report released by the Parents Resource ...
http://ericps.ed.uiuc.edu/npin/pnews/pnewn96/pnewn96f.html - size 4K - 23-May-97 - English

7. CEIDA Druglinks - Info Centre - PARENTS TALKING TO TEENAGERS ABOUT DRUGS

Query 1.2. Synonym expansion of teenager

(teenage* OR teen OR teens OR youth OR adolescent* OR kid* OR "high school")
AND drug*

About 249 documents match your query.

1. Adolescent Drug Abuse Treatment Outcome
Adolescent Drug Abuse Treatment Outcome. Executive Summary. This is a report on the evaluation of an inpatient adolescent drug abuse treatment program in..
http://www. cbc. med. umn. edu/~andy/drugabuse/adoltx. htm - size 3K - 28-Sep-96 - English

2. Poll finds parents overestimate communication with kids on drugs
03/03/97 - 07:26 PM ET - Click reload often for latest version. Poll finds parents overestimate communication with kids on drugs. NEW YORK - Most parents..
http://cgi.usatoday.com/elect/eq/eq17&htm - size 2K - 21-May-97 - English

3. Albany Youth Futures shows kids alternatives to drugs, alcohol/TITLE>
http://www *indreg.com/9-11-96/FEATURES/feature5.htm - size 5K - 13-Sep-96 - English*

4. IPRC Version - Keeping Youth Drug-Free - Exercise #3
Re-posted by the Indiana Prevention Resource Center at Indiana University Indiana's RADAR Network State Center. Exercise 3. Building Social Skills. Offer..
http://www. drugs. indiana. edu/pubs/radar/keeping/exer3. html - size 2K - 28-Jun-96 - English

5. Online NewsHour: Teen Drug Use Doubling -- August 20, 1996
THEY'RE NOT SAYING "NO" AUGUST 20, 1996. TRANSCRIPT. Two new and deeply troubling reports have just been released showing that drug abuse among 12 to 17...
http://web-crOl.pbs. org/newshour/bb/health/august96/teen_dru g~ab use_8-20. html - size 16K -10-Sep-96- English

6. Kmart: HOTNEWS/Kmart Kids Race Against Drugs Race Results
Kmart Kids Race Against Drugs. And the winner is... On Saturday. January 18. Jamie Barreiro of Port St. Lucie, FL, Joshua Brown of Willingboro, NJ and ...
http://www.kmart.com/hotnews/hotnews.stm. size 7K - 21-May-97 - English

11. OMH-RC Database Record: Drug Abuse Among Minority Youth: Methodological Issues Office of Minority Health Resource Center Database Record. When available, information on where these materials may be obtained has been listed below...
http://wwwomhrc.gov/mhr2/docs/95D2315.htm - size 3K - 1-May-97 - English

3.3.2 Thesauri – query results

Query 1.3. Plus synonym and hierarchic expansion of "drug*"

(teenage* OR teen OR teens OR youth* OR adolescent* OR kid* OR "high school")
AND (drug* OR substance* OR alcohol OR nicotine OR smoking OR cigarette*)
About 409 documents match your query.

1. Smoking is NOT for kids!
We believe smoking is for adults only. We therefore require that you be at least 18 years of age in order to view this site. Click below to enter the...
http://www.smokers.org/ - size 820 bytes - 20-Apr-97 - English

2. Adolescent Drug Abuse Treatment Outcome
Adolescent Drug Abuse Treatment Outcome. Executive Summary. This is a report on the evaluation of an inpatient adolescent drug abuse treatment program in..
http://www cbc. med. umn. edu/~andy/drugabuse/adoltx. htm - size 3K - 28-Sep-96 - English

3. Poll finds parents overestimate communication with kids on drugs
03/03/97 - 07:26 PM ET - Click reload often for latest version. Poll finds parents overestimate communication with kids on drugs. NEW YORK - Most parents..
http://cgi.usatodaycomielect/eq/eq] 7&htm - size 2K- 21-May-97 - English

4. Albany Youth Futures shows kids alternatives to drugs, alcohol/TITLE>
http../www. indregcoml9-11-96/FEATURESfeature5.htm - size 5K - 13-Sep-96 - English

5. IPRC Version - Keeping Youth Drug-Free - Exercise #3
Re-posted by the Indiana Prevention Resource Center at Indiana University Indiana's RADAR Network State Center. Exercise 3. Building Social Skills. Offer..
http://www. drugs. indiana. edu/pubs/radar/keeping/exer3. html - size 2K - 28-Jun-96 - English

6. Smoking still increasing among teens
Despite a chorus of ignorance one woman wanted to dance... To all of those people who say that national role models are a thing of the past, I want to...
http://www.bascchusgamma.org/bb_october/staff_view. html - size 5K - 11-Oct-96 - English

7. Online NewsHour: Teen Drug Use Doubling -- August 20, 1996
THEY'RE NOT SAYING "NO" AUGUST 20, 1996. TRANSCRIPT. Two new and deeply troubling reports have just been released showing that drug abuse among 12 to 17...
http://web-cr0l.pbs. org/newshour/bb/health/august96/teen_drug_abuse_8-20. html - size 16K -10-Sep-96- English

8. KCEOC SUBSTANCE ABUSE/YOUTH PROGRAM
KCEOC SUBSTANCE ABUSE/YOUTH PROGRAM. Address: 1611 First Street. Phone Number: *336-5310.* FAX Number: *336-5303.* Contact Person: Robert Cubit. Target Group..
http://www bakersfield. org/vdc/secondary/kceoc. html - size 2K - 15-Oct-96 - English

Query 1.4. Drug component more specific

(teenage* OR teen OR teens OR youth OR adolescent* OR kid* OR "high school")
AND (cocaine OR crack OR heroin)

2 documents match your query

1. Teenage "Huffing" - Worse Than Cocaine
Teenage "Huffing" - Worse Than Cocaine. May 22, 1996. MEEUWSEN: Imagine substances experts call deadlier than heroin or cocaine. Imagine that...
http://www.cbn.org/news/stories/huffinghtml - size 7K - 29-Oct-96 - English

2. Teen is arrested with a kil of crack cocaine
Teen is arrested with a kilo of crack cocaine. STROUDSBURG, Pa. (AP) - A 14-year-old New York City girl was busted during a bus trip through here last...
http..//www. recordernews. com/1 996/0703/natnews/teenare/teenare. html - size 2K - 25-May-97 English

3. Knowledge representation, Ontology & Metadata

3.4.1 Metadata

What is Metadata?

- ☑ Common definition: data about data, information added to a document or object to describe it
- ☑ Traditional library catalog is an example of metadata
- ☑ HTML “meta” tag – but it is abused sometimes

Metadata are needed for:

- ☑ Expand the description of the object
- ☑ Provide information of the data, such as where it can be found or what its uses might be
- ☑ Provide historical information which may be important to the organization holding the object
- ☑ Summarize some properties of the object
- ☑ Provide a standard description of the object

3.4.1 Metadata – Proprietary Definitions

PERSON	
ID	120
FIRSTNAME	Willy
LASTNAME	Bogner
BIRTHDAY	23
BIRTHMONTH	01
BIRTHYEAR	1942

IMAGEDATA	
ID	330976
TITLE	Olympic Wintergames 1960 in Squaw Valley
INFO	Willy Bogner in the slalom; minimum time in the first run
AUTHOR	Rübelt, Lothar
CREATION DATE	03-JUL-03
DATE	1960
FK_PERSON	120

IMAGEOBJECT	
ID	517849
INFO	http://www.bildarchivaustria.at/Bildarchiv//302/B1117424T4299954.jpg
MIMETYPE	image/jpeg
IMAGEWIDTH	2333
IMAGEHEIGHT	3147
FK_IMG_DATA	330976

Proprietary metadata describing a JPEG image

3.4.1 Metadata – Proprietary Definitions

```
<TVAMain "...">
  <ProgramDescription>
    <ProgramInformationTable>
      <ProgramInformation programId="crid://bbc.co.uk/123456789">
        <BasicDescription>
          <Title>Lake Placid 1980, Alpine Skiing, I. Stenmark</Title>
          <Synopsis>Ingmar Stenmark's (SWE-Alpine skiing) victory in
            the Giant Slalom in Lake Placid</Synopsis>
          <Genre href="urn:tva:metadata:cs:ContentCS:2004:3.1.1.9">
            <Name>Sports</Name>
          </Genre>
          <CreditsList>
            <CreditsItem role="urn:mpeg:mpeg7:cs:RoleCS:2001:AUTHOR">
              <PersonName>
                <mpeg7:GivenName>John</mpeg7:GivenName>
                <mpeg7:FamilyName>Doe</mpeg7:FamilyName>
              </PersonName>
            </CreditsItem>
          </CreditsList>
          <CreationCoordinates>
            <CreationLocation>us</CreationLocation>
          </CreationCoordinates>
        </BasicDescription>
      </ProgramInformation>
    </ProgramInformationTable>
  </ProgramDescription>
</TVAMain>
```

TV-Anytime metadata describing a video

3.4.1 Metadata- Dublin Core Standard

The Dublin Core: Metadata for Digital Libraries

- The Dublin Core is a set of simple name/value properties that can describe online resources.
  - ☑ Usually for web content but generally usable
  - ☑ Intended to help classify and search online resources.
- Initial set defined by 1995 Dublin, Ohio meeting.
- DC elements may be either embedded in the data or in a separate repository.
- 15 Elements of Dublin Core:

1. Contributor
2. Coverage
3. Creator
4. Date
5. Description
6. Format
7. Identifier
8. Language
9. Publisher
10. Relation
11. Rights
12. Source
13. Subject
14. Title
15. Type

3.4.1 Metadata- Dublin Core

Elements of Dublin Core

- Content elements:

  coverage, description, type, relation, source, subject, title.

- Intellectual property elements:

  contributor, creator, publisher, rights

- Instantiation elements:

  date, format, identifier, language

3.4.1 Metadata- Dublin Core

contributor	An entity responsible for making contributions to the resource.
coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
creator	An entity primarily responsible for making the resource.
date	A point or period of time associated with an event in the lifecycle of the resource.
description	An account of the resource.
format	The file format, physical medium, or dimensions of the resource.
identifier	An unambiguous reference to the resource within a given context.
language	A language of the resource.
publisher	An entity responsible for making the resource available.
relation	A related resource.
rights	Information about rights held in and over the resource.
source	A related resource from which the described resource is derived.
subject	The topic of the resource.
title	A name given to the resource.
type	The nature or genre of the resource.

3.4.1 Metadata- Dublin Core

An example set of Dublin Core elements

```
<head profile="http://dublincore.org">
<title> ... </title>

<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />

<meta name="DC.Identifier" schema="DCterms:URI"
        content="http://tutorialsonline.info/Common/DublinCore.html" />
<meta name="DC.Format" schema="DCterms:IMT" content="text/html" /> <meta name="DC.Title" xml:lang="EN" content="Dublin Core Tutorial" />
<meta name="DC.Creator" content="Alan Kelsey" />
<meta name="DC.Subject" xml:lang="EN" content="Dublin Core Meta Tags" />
<meta name="DC.Publisher" content="Alan Kelsey, Ltd." />
<meta name="DC.Publisher.Address" content="alan@tutorialsonline.info" />
<meta name="DC.Contributor" content="Alan Kelsey" />
<meta name="DC.Date" scheme="ISO8601" content="2007-01-06" />
<meta name="DC.Type" content="text/html" />
<meta name="DC.Description" xml:lang="EN"
        content="Learning Advanced Web Design can be fun and easy! Look at a site designed specifically to help you learn how to design web pages
                with proper tags, styles, and scripting." />
<meta name="DC.Identifier" content="http://tutorialsonline.info/Common/DublinCore.html" />
<meta name="DC.Relation" content="TutorialOnline.info" scheme="IsPartOf" />
<meta name="DC.Coverage" content="Hennepin Technical College" />
<meta name="DC.Rights" content="Copyright 2011, Alan Kelsey, Ltd.  All rights reserved." />
<meta name="DC.Date.X-MetadataLastModified" scheme="ISO8601" content="2007-01-06" />
<meta name="DC.Language" scheme="dcterms:RFC1766" content="EN" />
```

http://www.tutorialsonline.info/Common/DublinCore.html

3.4.1 Metadata- Dublin Core

Dublin Core metadata describing an image with OAI-PMH Interface

```
<OAI-PMH "...">
    ...
    <metadata>
        <oai_dc:dc "...">
            <dc:title>Sydney Olympics 2000, marathon runners cross Sydney
                Harbour Bridge [picture] /</dc:title>
            <dc:creator>Mahony, David (David James)</dc:creator>
            <dc:format>1 photograph : gelatin silver ; image 26.9 x 38.4 cm.
                on sheet 30.5 x 40.3 cm.</dc:format>
            <dc:coverage>New South Wales</dc:coverage>
            <dc:date>2000</dc:date>
            <dc:description>Photograph by David Mahony -- On reverse in pencil.;
                Condition: Good. Group of [marathon] runners feature
                eventual Gold Medal Winner Gezahgne Abero of Ethiopia (No.
                1651) [Sydney, N.S.W., September 2000]</dc:description>
            <dc:subject>Runners (Sports) -- Australia -- Portraits.</dc:subject>
            <dc:subject>Sydney Harbour Bridge (Sydney, N.S.W.)</dc:subject>
            <dc:subject>Olympic Games (27th :, 2000 : Sydney, N.S.W.)</dc:subject>
            <dc:subject>Marathon running -- Australia -- Photographs.</dc:subject>
            <dc:subject>Sportsmen and sportswomen.</dc:subject>
            <dc:type>Image</dc:type>
            <dc:identifier>nla.pic-an22842546</dc:identifier>
            <dc:source>Item held by National Library of Australia</dc:source>
            <dc:rights>You may save or print this image for research and study.</dc:rights>
            <dc:identifier>http://nla.gov.au/nla.pic-an22842546</dc:identifier>
        </oai_dc:dc>
    </metadata>
    ...
</OAI-PMH>
```

3.4.1 Metadata- Dublin Core

Element Refinements

- Many of these, and extensible
- See http://dublincore.org/documents/dcmi-terms/ for the comprehensive list of elements and refinements

Encoding: DC elements are independent of the encoding syntax

- Rules exist to map the DC into
  - HTML
  - RDF/XML

3.4.1 Metadata- Trends and Lessons

- Online data repositories becoming increasingly important, so need pedigree, preservation, maintenance, and annotation.
- XML is the method of choice for describing metadata.
  - Extend “tagging” so that the **user** could know what a data item means, e.g.,

```
<temperature>75</temperature>
<time-interval>75</time-interval>
<price><euro>75</euro></price>
```

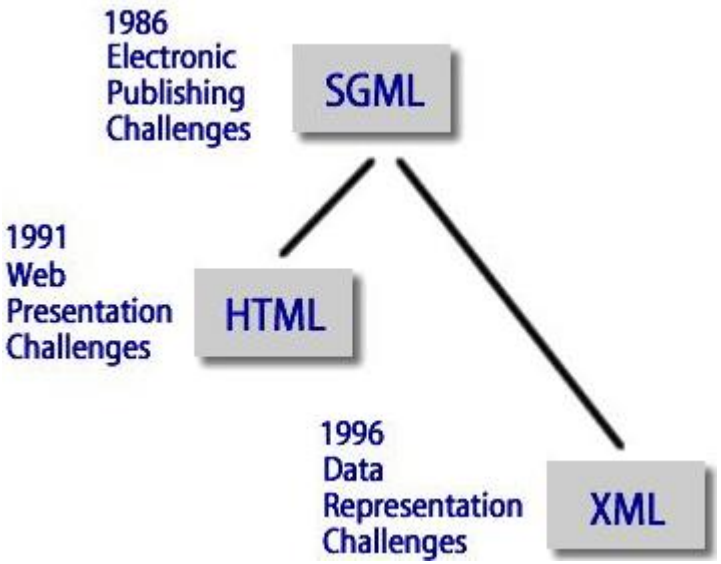
  - Provides **syntax** rules but does not really encode meaning (**semantics**)

- Semantic Web and Linked Data: in addition to tagging the data, agreement on the tags to be used needs to be made

3.4.2 XML

XML - eXtensible Markup Language

- Based on Standard Generalized Markup Language (SGML)
- Version 1.0 introduced by World Wide Web Consortium (W3C) in 1998, Version 2.0, 2006
- Vehicle for data exchange on the Web



e**X**tensible **M**arkup **L**anguage (XML)

- **XML** 1.0 Specification defined by **W3C** since 1998

- **XML** := designed to **transport** data
- **HTML** := designed to **display** data

- → XML is a software and hardware **independent** language for carrying data

- XML is *very* important for the *Web* (but also other applications) to exchange data

e**X**tensible **M**arkup **L**anguage (XML)

- XML tags are ***not*** predefined
- ***BUT***: XML tags can be formally defined

- Definition of new tags in:
  - **XML Schema**
  - **Document Type Definition (DTD)**

- Meta-language for defining new *languages / models*

XML Based Languages

- Graphic
  - SVG
  - Collada
- Text
  - Office Open XML (docx)
  - XHTML
- Geodata
  - Geography Markup Language (GML)
  - Keyhole Markup Language (KML)
  - OpenStreetMap Format (OSM)
  - City Geography Markup Language (CityGML)
- Communication
  - Chat systems (z.B. XMPP)
  - Web Feeds (z.B. RSS)
  - Web Service Communication (SOAP, WSDL, ...)

XML Example: Bookstore

```
<bookstore>
  <book category="CHILDREN">
    <title>Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>
  <book category="WEB">
    <title>Learning XML</title>
    <author>Erik T. Ray</author>
    <year>2003</year>
    <price>39.95</price>
  </book>
</bookstore>
```

Root Element
(*parent* of all other elements)

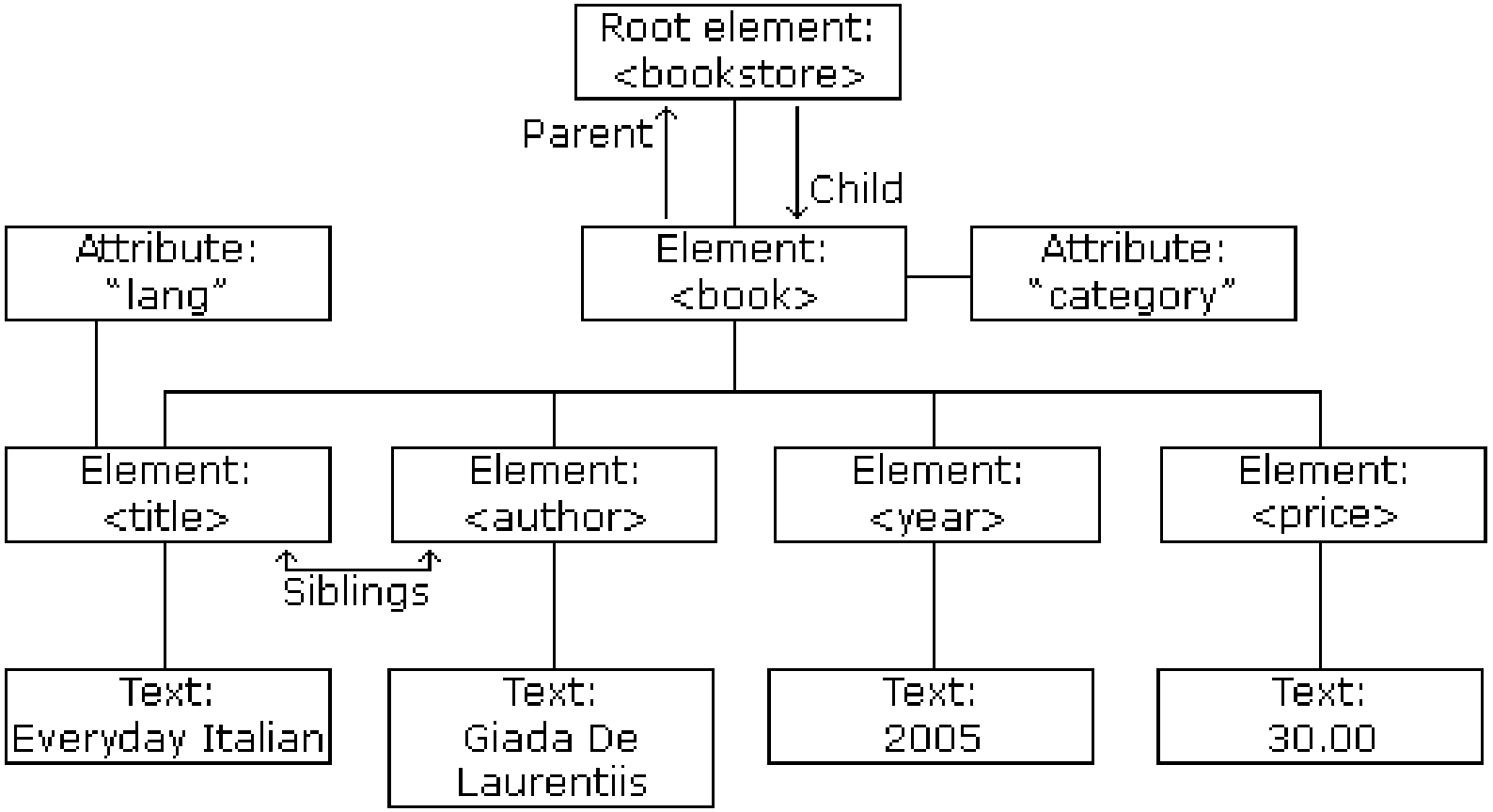
Attribute

Element Value

Child Element

XML Example: Bookstore

- All XML documents form a tree structure:



Well-formed XML Documents…

- …must have correct syntax:
  - XML document must have ONE root
  - All elements must have closing tag:
    - `<book> … </book>`
  - Or: element is empty:
    - **<**`book category="CHILDREN"` **/>**
  - XML tags are case sensitive
    - `<`**b**`ook> … </`**B**`ook>` => incorrect
  - …
  - (NOTE: all this is not the case for HTML documents!)

Valid XML Documents…

- …must be well-formed **and** conform to an ***XML Schema*** (or DTD)

- (Note: XML Schema is successor of DTDs)

XML Schema

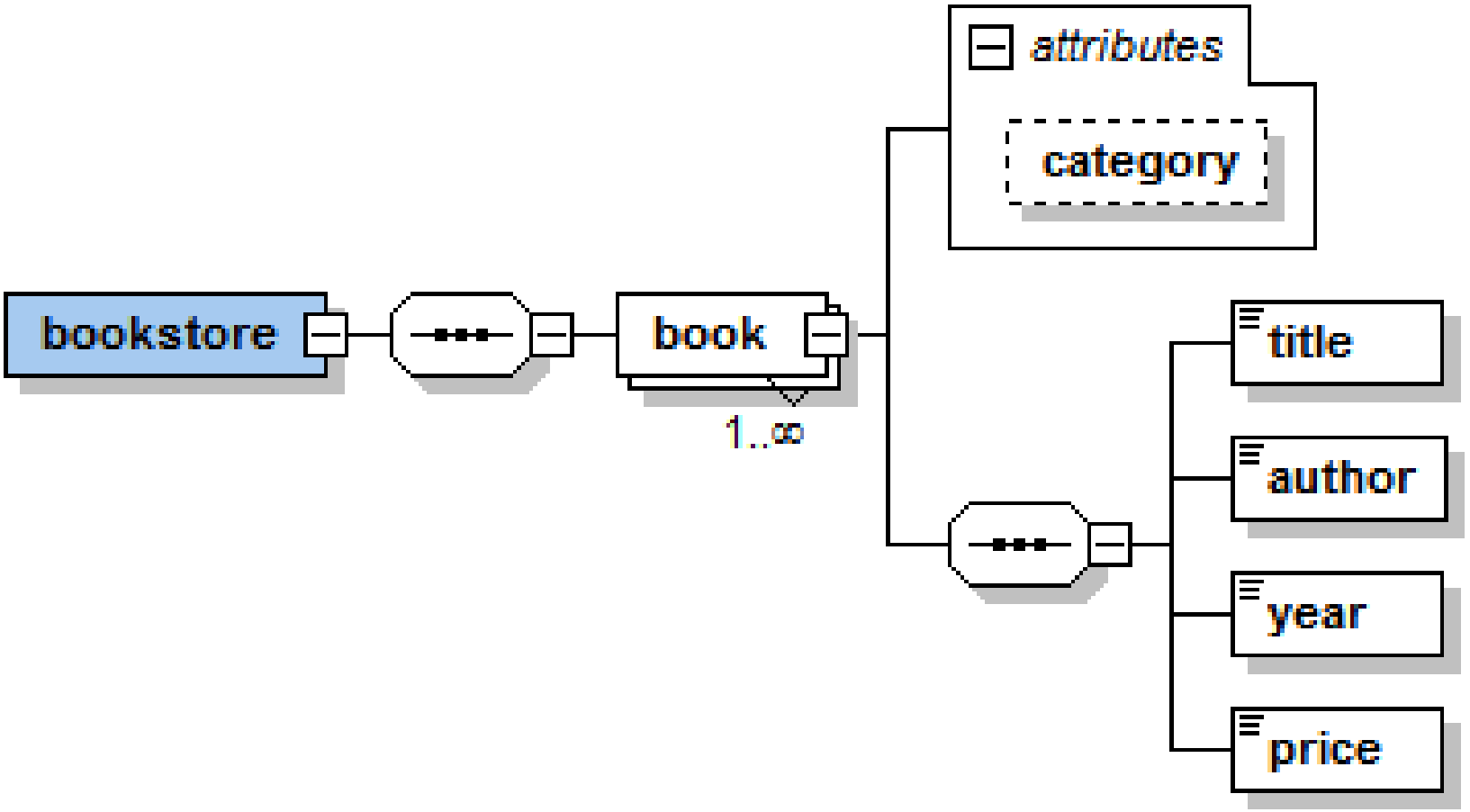
- Describes structure of XML documents (= *instances*)
- Stored in XML Schema Definition (XSD) file

→ can be used to formally define new *languages / models*

  - (e.g., GML, KML, SVG etc.)

- Technical Functionalities:
  - Definition of elements & attributes that can appear in XML
  - Definition of XML tree structure
  - Definition of data types for elements & attributes
  - Definition of default values for element & attributes
  - ...

XML Schema Example: Bookstore



```
<xs:schema>
   <xs:element name="bookstore">
      <xs:complexType>
         <xs:sequence>
            <xs:element name="book" maxOccurs="unbounded">
               <xs:complexType>
                  <xs:sequence>
                     <xs:element name="title"  type="xs:string" />
                     <xs:element name="author" type="xs:string" />
                     <xs:element name="year"   type="xs:short"  />
                     <xs:element name="price"  type="xs:double" />
                  </xs:sequence>
                  <xs:attribute name="category" type="xs:string"/>
               </xs:complexType>
            </xs:element>
         </xs:sequence>
      </xs:complexType>
   </xs:element>
</xs:schema>
```

XML + XML Schema Example: Bookstore

```
<?xml version="1.0" encoding="UTF-8"?>
<UKON:bookstore
  xmlns:UKON="http://www.uni-konstanz.de"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.uni-konstanz.de
xsdExample.xsd">

  <book category="CHILDREN">
    <title>Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>

</UKON:bookstore>
```

Namespace Definition

Schema Reference

3.4.3 RDF - representing information about resources

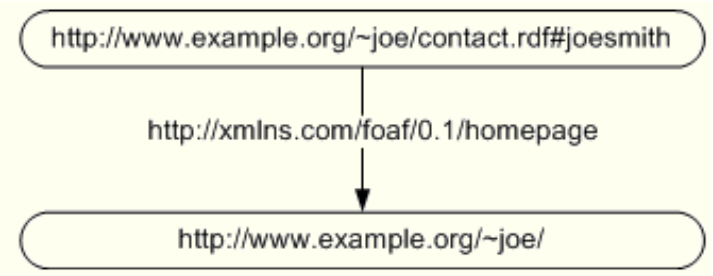
RDF - Resource Description Framework

- A framework for representing information about resources (such as the title, author, modification date, content, and copyright information) in a graph form.
- Primarily intended for representing metadata about *WWW* resources
- A standard model for data interchange on the Web
- The RDF language namespace prefix is usually rdf and is (syntactically) defined at http://www.w3.org/1999/02/22-rdf-syntax-ns#

Benefit: One can parse the semantic tree, which end up giving him/her a set of (possibly mutually referential) triples and then he/she can use the ones he/she wants and ignoring the ones he/she don't understand.

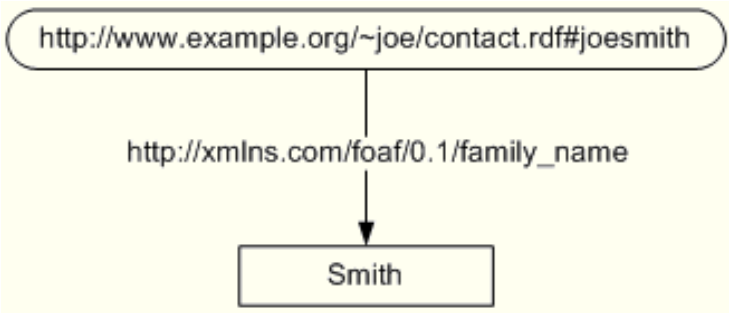
3.4.3 RDF

- Information is represented by triples *subject-predicate-object* in RDF



Joe Smith
|
homepage identified by
|
http://www.example.org/~joe/

- Object can be also a “literal” - a constant string value



Joe Smith
|
has surname
|
Smith

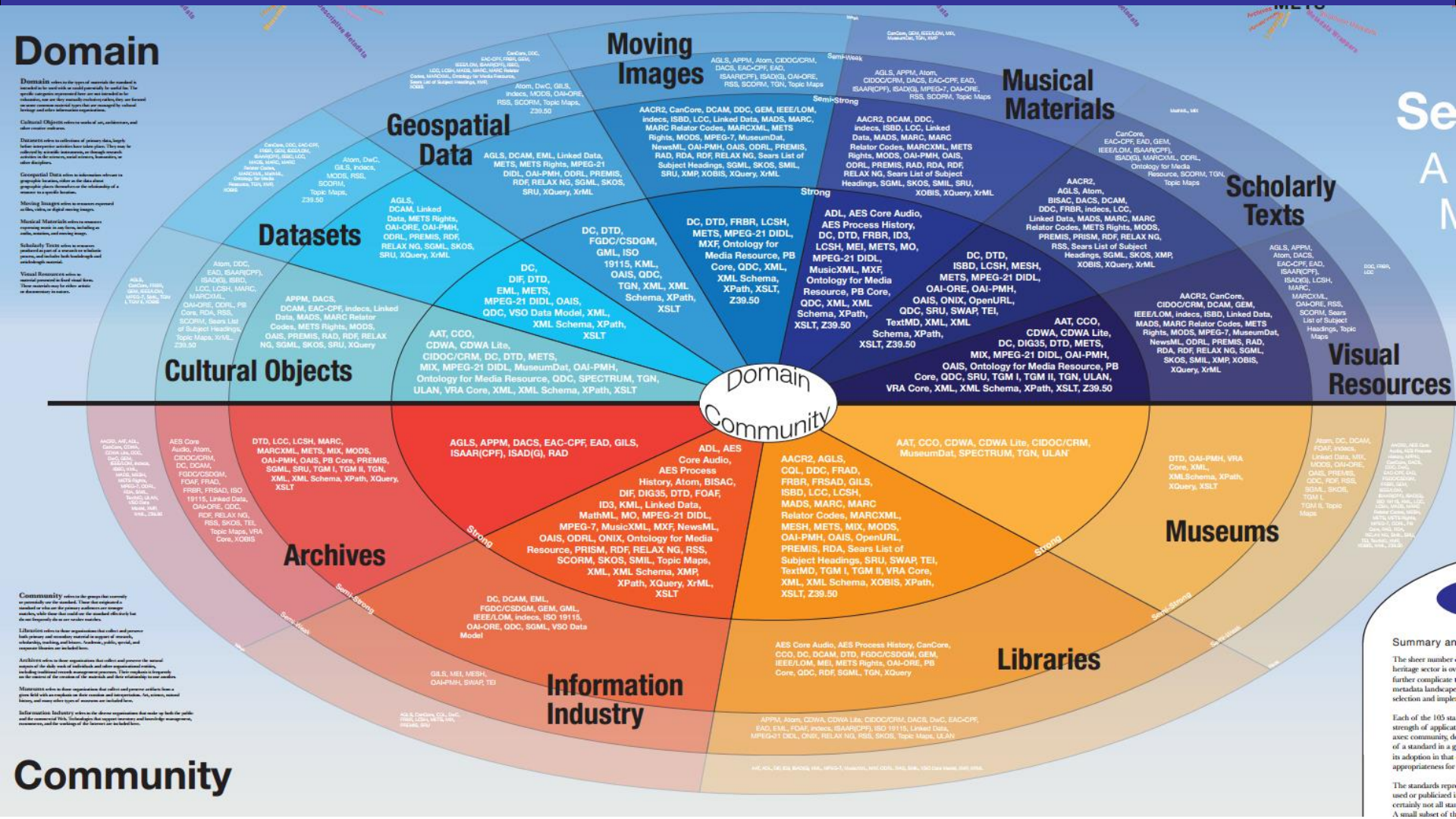
3.4.3 Semantic Web

- Traditionally, database are isolated items, and each one was interpreted by specific queries and software just for that database
- The **semantic web** tend to be general so that everyone can write agents that visit semantic web pages and work with them.
  - provides general languages for describing any metadata
  - enable knowledge representation and limited machine reasoning.

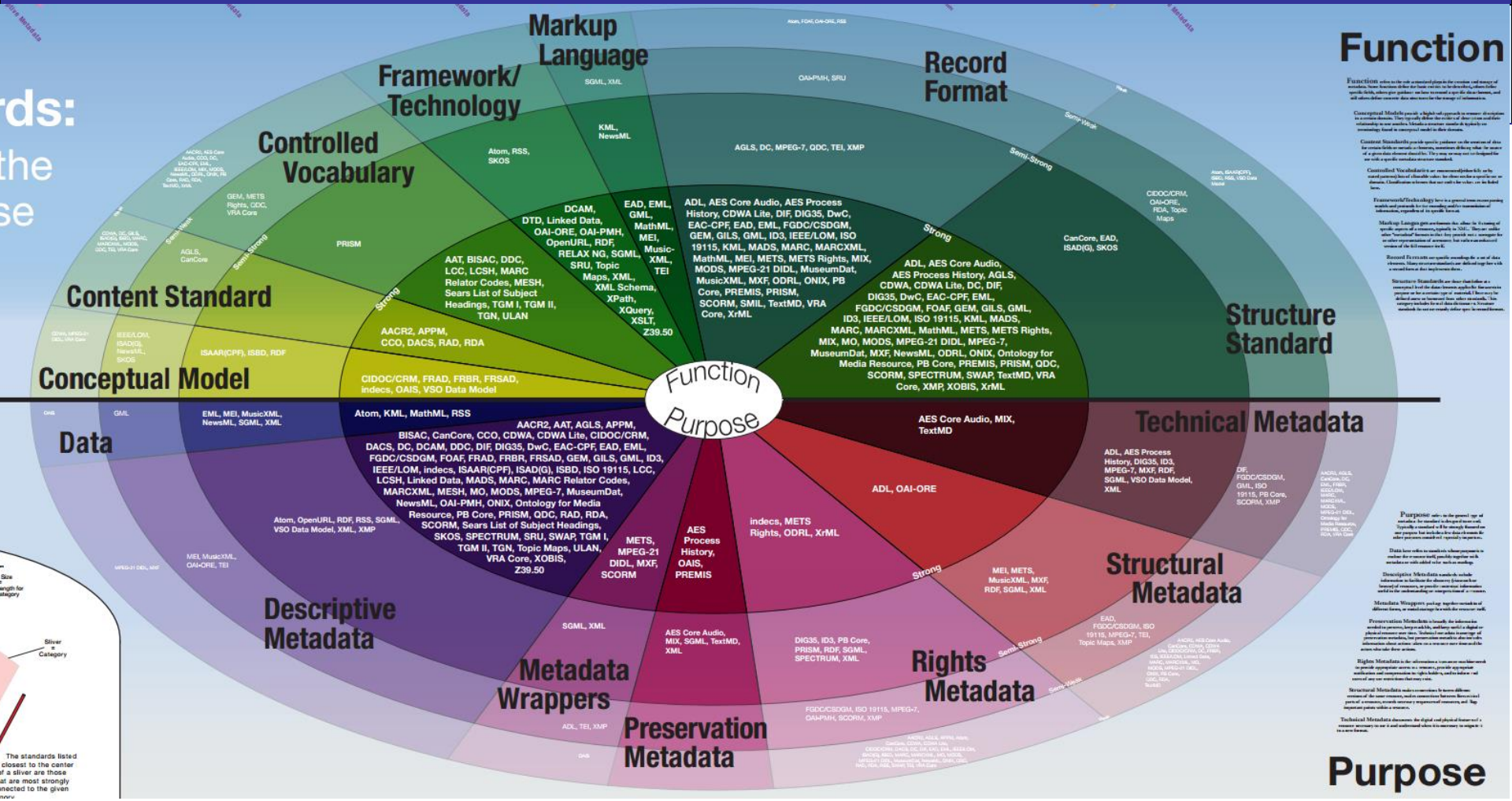
3.4.3 Semantic Web

- **Semantic Web** is an attempt to merge the database area with the web, so that data schemas and data files can be stored on the webpages and accessed by programmers
  - Each user has to post a “data schema” for their data
  - Each user has to define each data element in terms of that schema
  - Each user has to define the data elements in a *controlled vocabulary*
  - Each user has to define the data relationships in a *controlled vocabulary*
- **Controlled vocabulary** here indicates an ontology, giving precise definitions of the actual items to be used.
- It is not clear yet how these ontologies are going to be created or defined.

  Open questions: Will people all cooperate? Will people try to do such labeling well?



Seeing Standards:
A Visualization of the Metadata Universe
http://www.dlib.indiana.edu/~jenlrile/metadatamap/



Seeing Standards:
A Visualization of the Metadata Universe
http://www.dlib.indiana.edu/~jenlrile/metadatamap/

Linked Data

- Part of the Semantic Web:

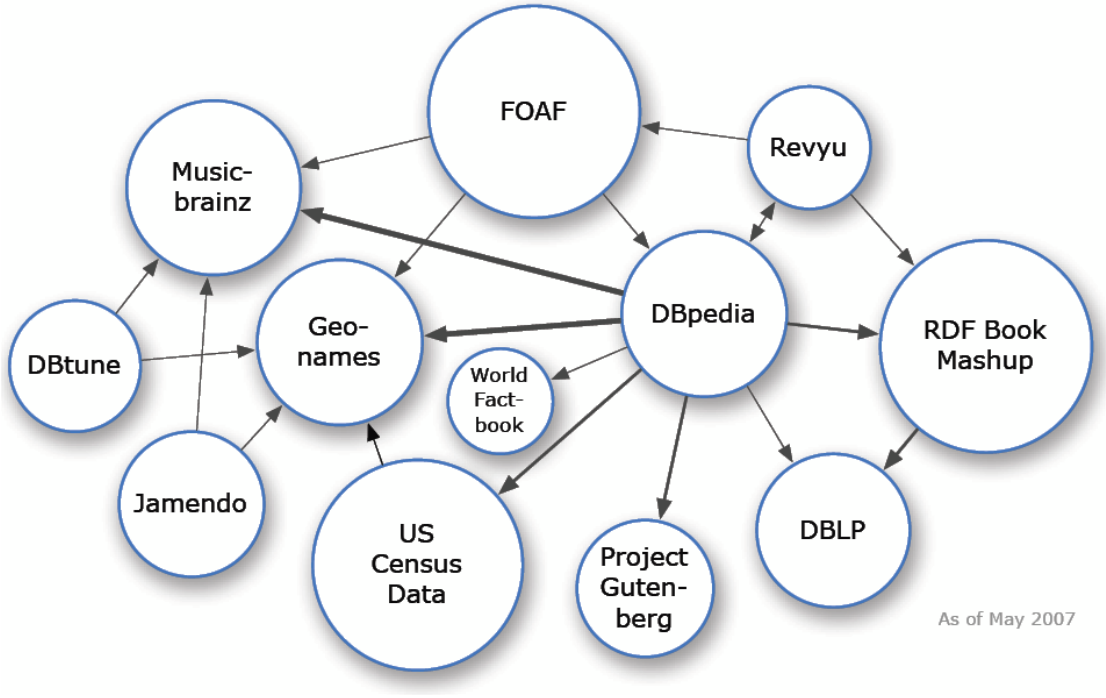
Linked Data

– Interlinked RDF resources on the Web

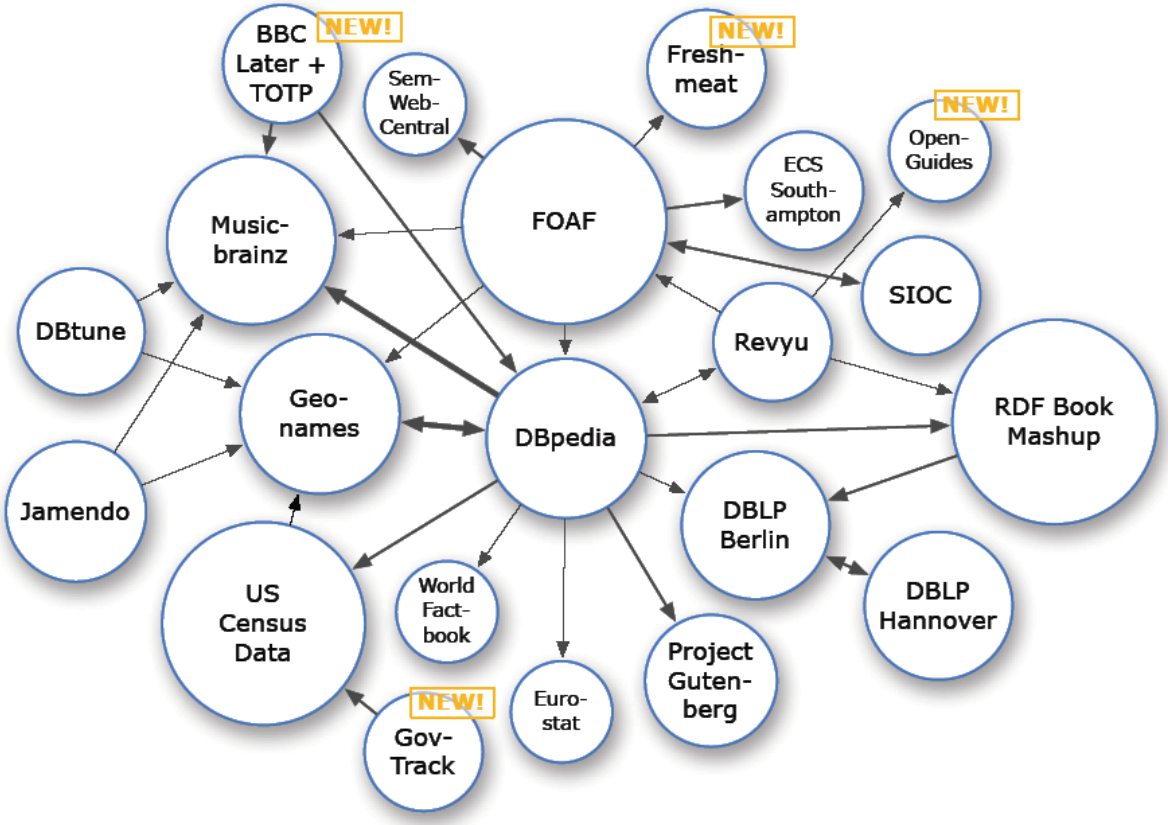
– A more recent research & development topic which is getting more and more attention
– E.g., see the TED talk of Sir Tim Berners-Lee

Linked Data Cloud – March 2007

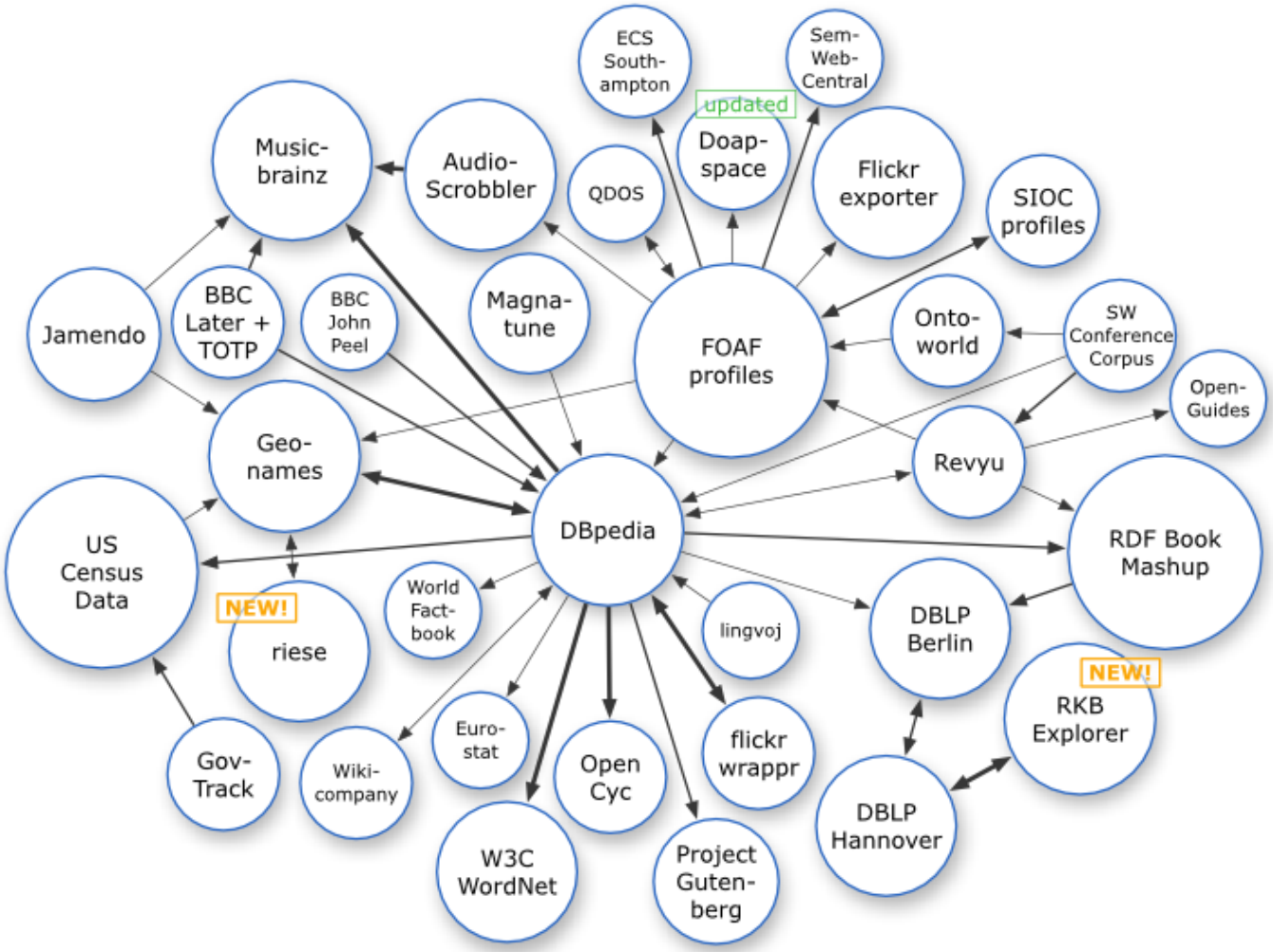
- Already available linked data:
  - people, companies, publications, books, movies, music, television programs, genes, proteins…



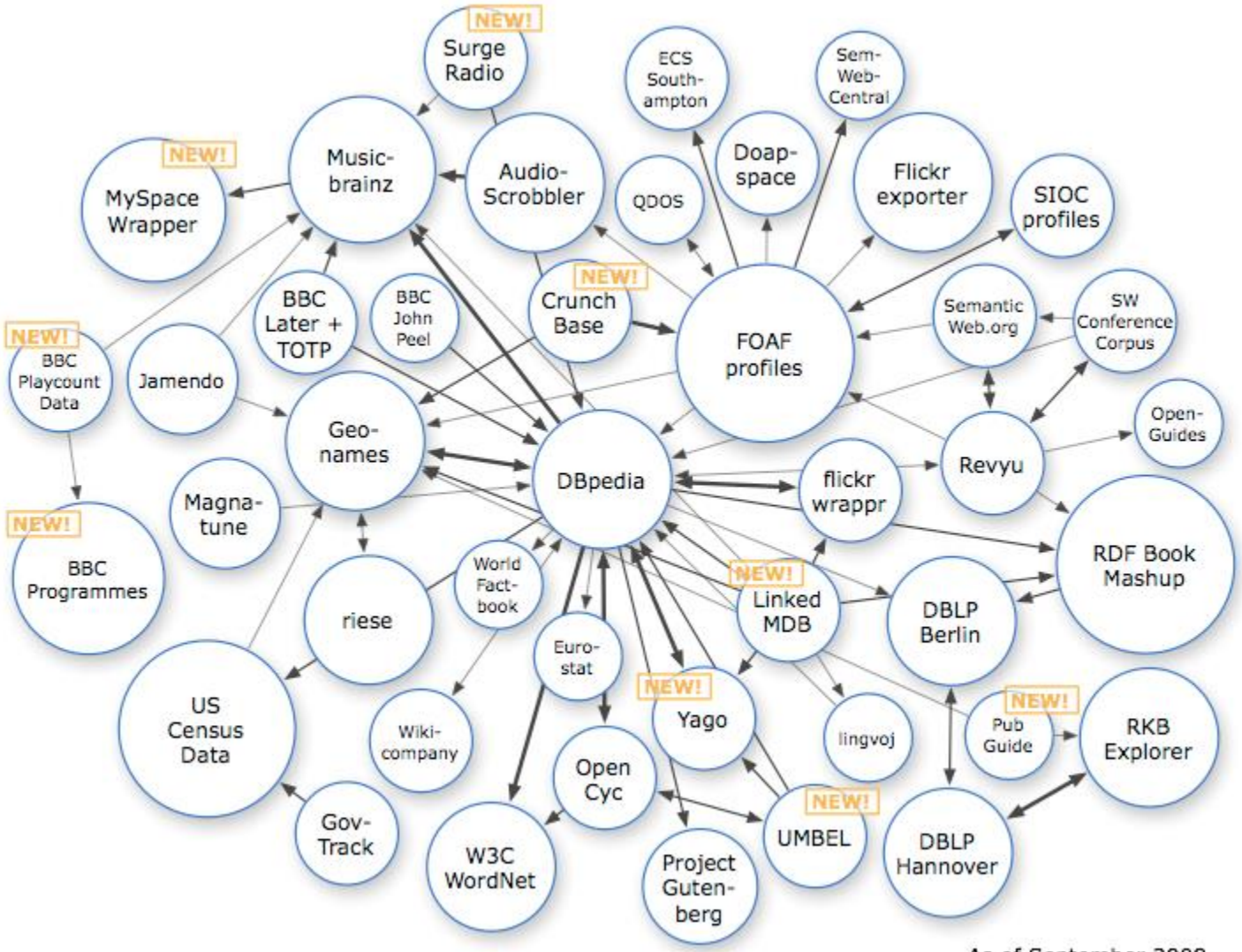
Linked Data Cloud – August 2007



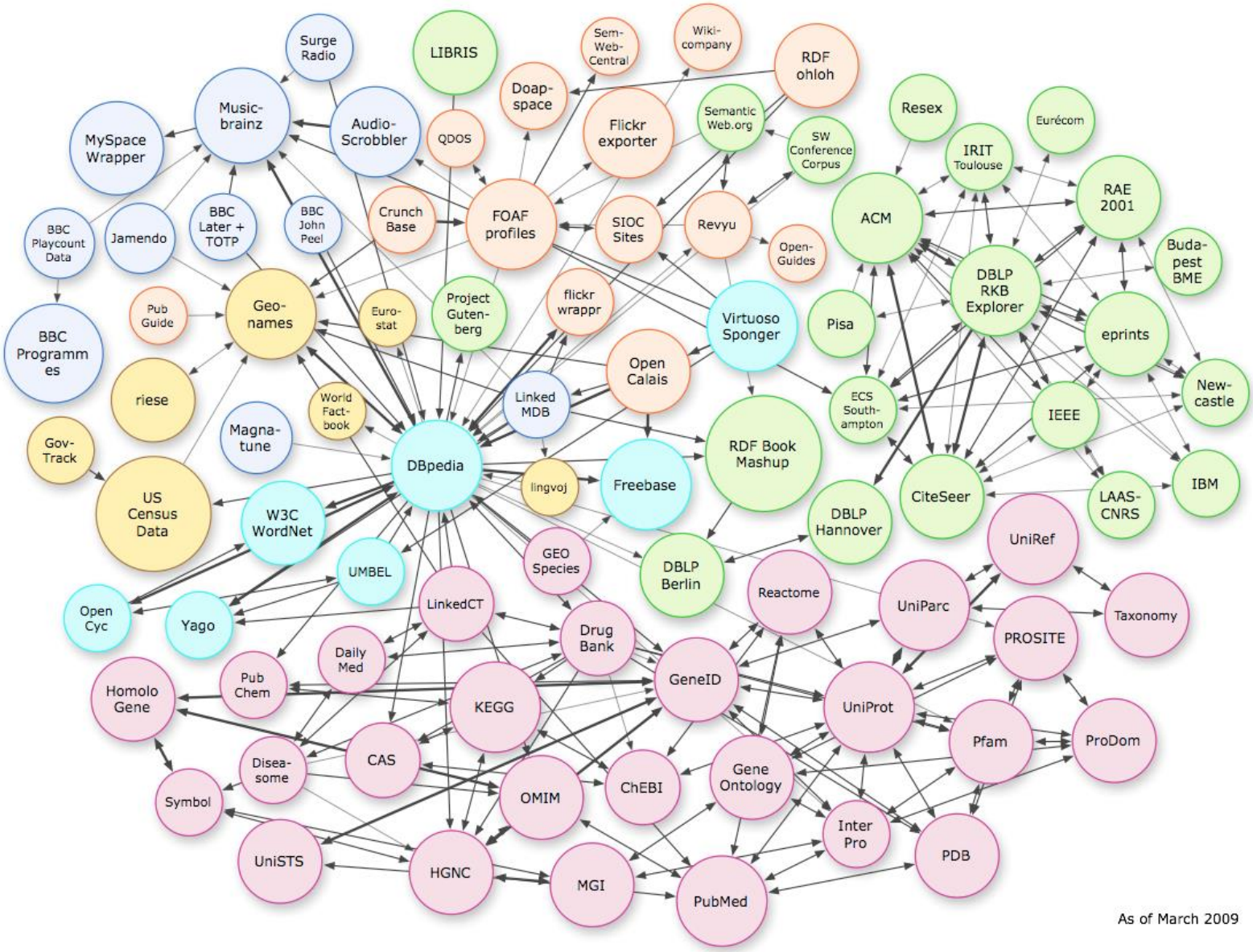
Linked Data Cloud – March 2008



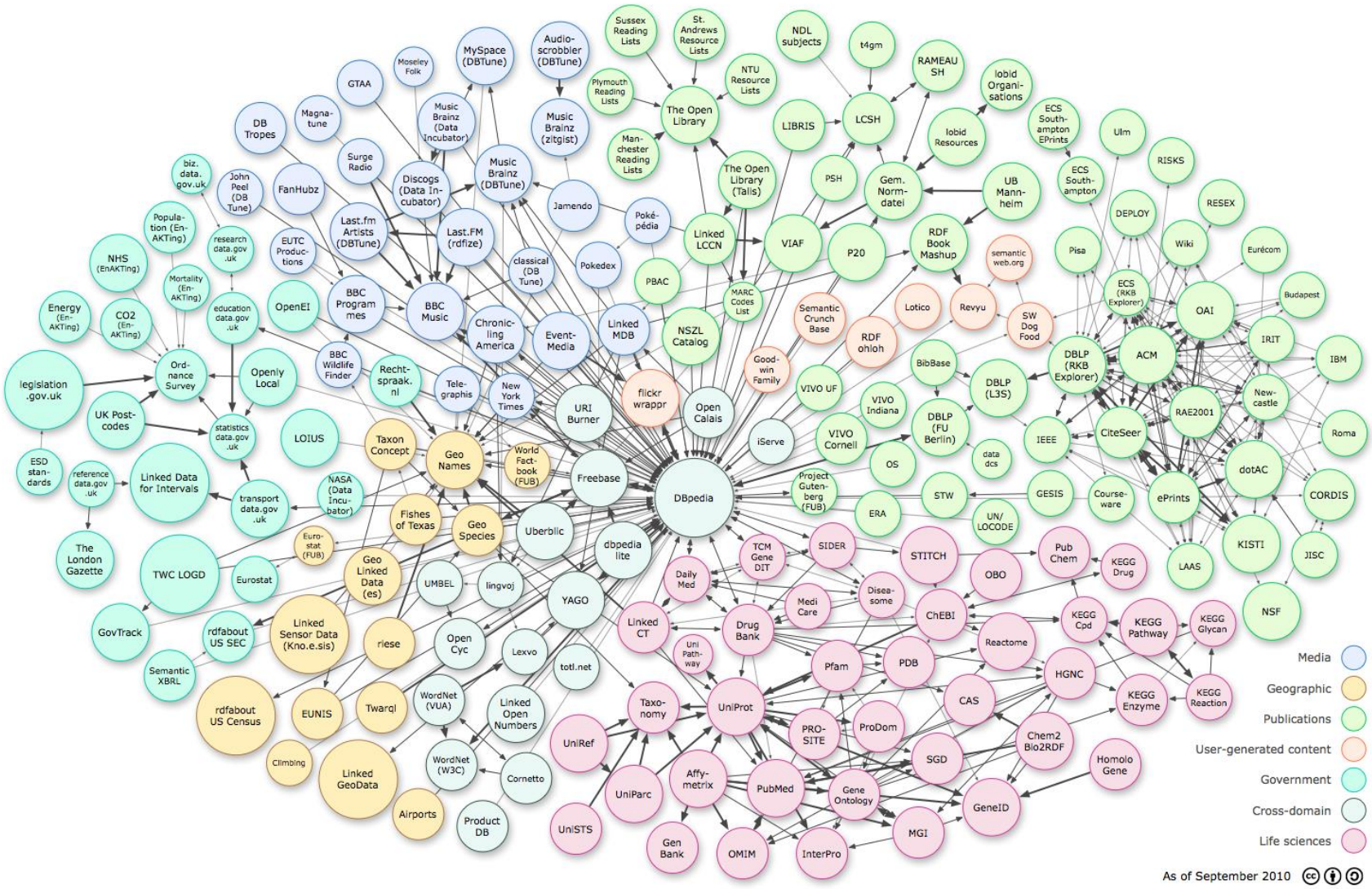
Linked Data Cloud – September 2008



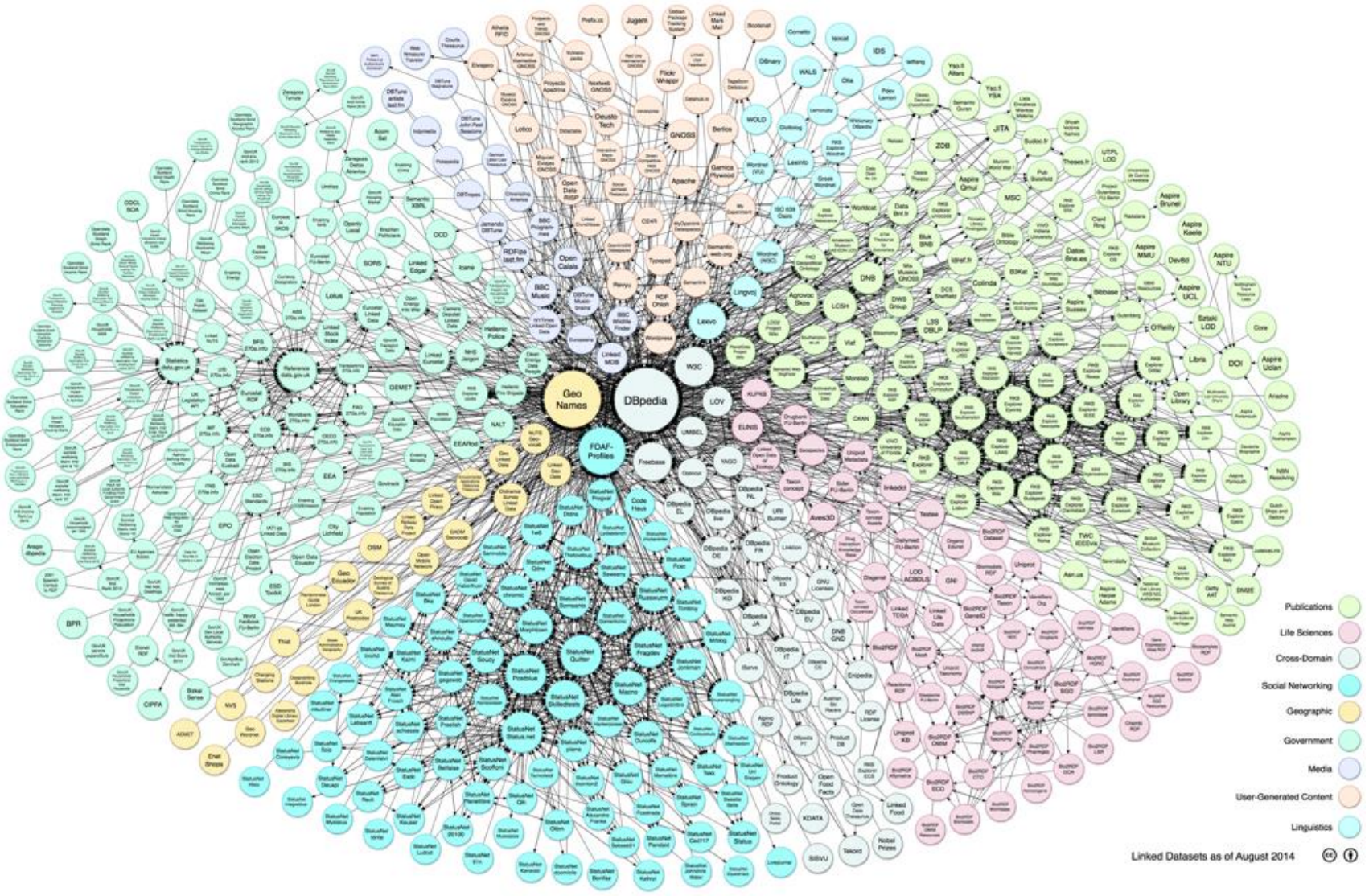
Linked Data Cloud – March 2009



Linked Data Cloud – September 2010



Today


Geo Names
DBpedia
FOAF-Profiles
W3C
LOV
Publications
Life Sciences
Cross-Domain
Social Networking
Geographic
Government
Media
User-Generated Content
Linguistics
Linked Datasets as of August 2014

3. Knowledge representation, Ontology & Metadata

3.5 The Vector Space Model

- Document represented by a vector of terms
  - Words (or word stems)
  - Phrases (e.g. computer science)
  - Removes words on “stop list”
- Often assumed that terms are uncorrelated.
- Correlations between term vectors implies a similarity between documents.
- For efficiency, an inverted index of terms is often stored.

3.5 The Vector Space Model

Document Vectors

- Documents are represented as “bags of words”
- Represented as vectors when used computationally
  - A vector here often is an array of floats
  - Has direction and magnitude
  - Each vector holds a place for every term in the collection
  - Therefore, most vectors are sparse

3.5 The Vector Space Model

Vector Representation

- Documents and Queries are represented as vectors.
- Position 1 corresponds to term 1, position 2 to term 2, position *t* to term *t*

$$D_i = w_{d\,i1}, w_{d\,i2}, ..., w_{d\,it}$$

$$Q = w_{q1}, w_{q2,}, ..., w_{qt}$$

$$w = 0 \text{ if a term is absent}$$

- Weight terms high if they are
  - frequent in relevant documents … BUT
  - infrequent in the collection as a whole

3.5 The Vector Space Model

What values to use for terms

- Simple Model: Boolean (term present /absent)

- Improved model: TF*IDF
  - TF (term frequency) - Count of times term occurs in document.
    - The more times a term *t* occurs in document *d* the more likely it is that *t* is relevant to the document.
    - Used alone, favors common words, long documents.
  - DF document frequency
    - The more a term *t* occurs throughout all documents, the more poorly *t* discriminates between documents
  - TF-IDF term frequency * inverse document frequency -
    - High value indicates that the word occurs more often in this document than average.

Document Vectors

Document ids

	nova	galaxy	heat	h'wood	film	role	diet	fur
A	1.0	0.5	0.3					
B	0.5	1.0						
C				1.0	0.8	0.7		
D				0.9	1.0	0.5		
E							1.0	1.0
F							0.9	1.0
G	0.5		0.7			0.9		
H		0.6	1.0	0.3	0.2	0.8		
I				0.7	0.5		0.1	0.3

3.5 The Vector Space Model

What to Evaluate?

- What can be measured that reflects users' ability to use system? (Cleverdon 66)
  - Coverage of Information
  - Form of Presentation
  - Effort required/Ease of Use
  - Time and Space Efficiency
  - Recall
    proportion of relevant material actually retrieved
  - Precision
    proportion of retrieved material actually relevant

effectiveness

3.5 The Vector Space Model

Relevance

- In what ways can a document be relevant to a query?
  - Answer precise question precisely.
  - Partially answer question.
  - Suggest a source for more information.
  - Give background information.
  - Remind the user of other knowledge.
  - Others ...

3.5 The Vector Space Model

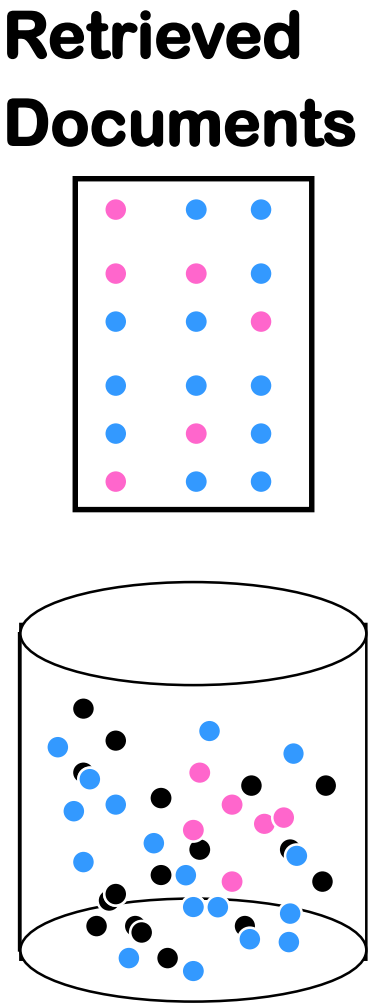
Standard IR Evaluation

- Precision

$$\frac{\#\text{ relevant retrieved}}{\#\text{ retrieved}}$$

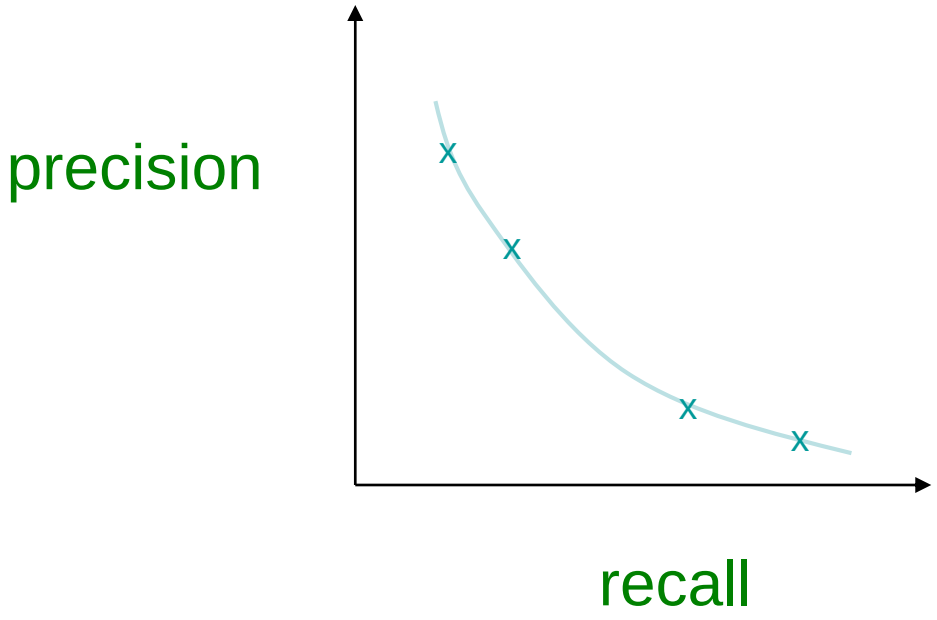
- Recall

$$\frac{\#\text{ relevant retrieved}}{\#\text{ relevant in collection}}$$



Precision/Recall Curves

- There is a tradeoff between Precision and Recall
- So measure Precision at different levels of Recall



3.5 The Vector Space Model

The E-Measure

Combine Precision and Recall into one number (van Rijsbergen 79)

$$E = 1 - \frac{b^2 PR + PR}{b^2 P + R}$$

P = precision
R = recall
b = measure of relative importance of P or R

For example,
b = 0.5 means user is twice as interested in precision than in recall

Summary

- Common problem with DL searching
- Classifications & Ontologies

  definition, components, & examples
- Indexing: words and thesauri

  words, thesauri and query expansion
- Metadata & Semantic Web

  functionality, examples, and evaluation of different standards

  definition, motivation of semantic web
- Vector models

  use, evaluation

Learning Goals

- Understand the basic problem of organization and retrieval in text documents
- Know basic classification schemes
- Know and understand main data structures to describe content and knowledge (Dictionary, Metadata, Thesaurus, Ontology)
- Understand how to apply them for automatic text processing and query expansion
- Understand the basic tf*idf vector model for text similarity and precision/recall evaluation for retrieval