Semantic Technologies

(Knowledge Graphs and All That)

Michael Zakharyaschev

Department of Computer Science and Information Systems

Birkbeck, University of London

- email: zmishaz@gmail.com
- homepage: http://www.dcs.bbk.ac.uk/~michael
- ST Web page: http://www.dcs.bbk.ac.uk/~michael/sw15/sw15.html

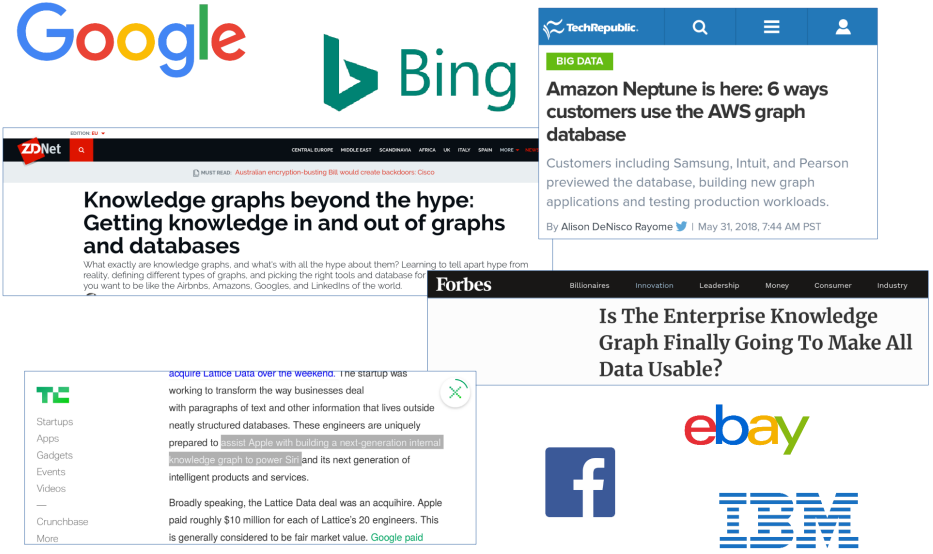
Acknowledgements

When working on this module, I used some materials developed by

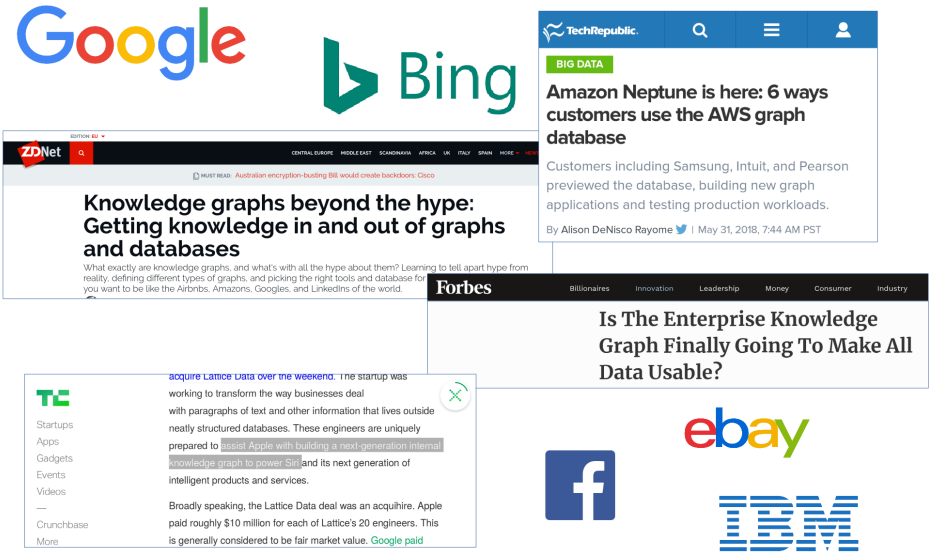
- Prof. Dr. Markus Krötzsch (Dresden)
  https://iccl.inf.tu-dresden.de/web/Knowledge_Graphs_(WS2018/19)
- Prof. Dr. Jens Lehmann (Leipzig)
  https://sewiki.iai.uni-bonn.de/teaching/lectures/kga/2018/start
- Prof. Dr. Sebastian Rudolph (Dresden)
  https://iccl.inf.tu-dresden.de/web/Foundations_of_Semantic_Web_Technologies_(SS2017)/en
- Prof. Martin Giese (Oslo)

Special thanks are due to Prof Frank Wolter (Liverpool),
Dr Stanislav Kikot (Oxford) and Dr Roman Kontchakov (London)

Knowledge Graphs are everywhere



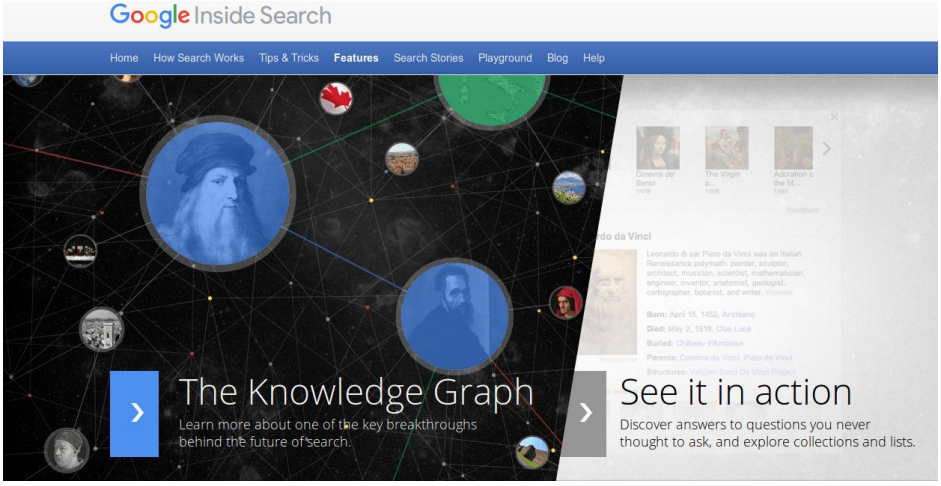
Knowledge Graphs are everywhere



What is a knowledge graph?

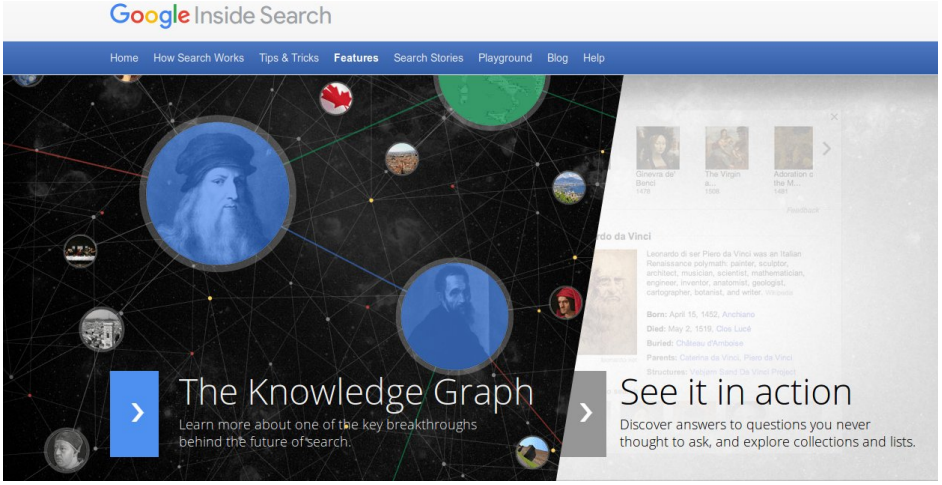
Google: "... we have been working on an intelligent model — in geek-speak, a 'graph' — that understands real-world entities and their relationships to one another: things, not strings."

Google Knowledge Graph (2012)



Things, not Strings!

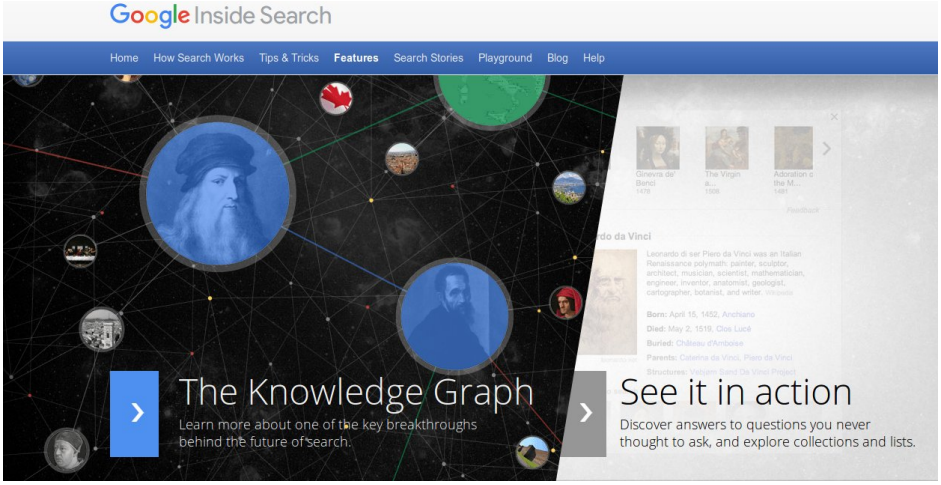
Google Knowledge Graph (2012)



Things, not Strings!

- Google's Knowledge Vault
- Yahoo!'s Knowledge Graph
- Microsoft's Bing Satori
- Facebook's Entities Graph
- LinkedIn knowledge graph
- Wikidata
- DBpedia
- YAGO
- Amazon Neptune
- Apple also working ...

Google Knowledge Graph (2012)



Things, not Strings!

- Google's Knowledge Vault
- Yahoo!'s Knowledge Graph
- Microsoft's Bing Satori
- Facebook's Entities Graph
- LinkedIn knowledge graph
- Wikidata
- DBpedia
- YAGO
- Amazon Neptune
- Apple also working ...

Exercise: Represent the information on page 1 as a 'knowledge graph'

So, what is a Knowledge Graph?

"...major companies such as Google, Yahoo!, Microsoft, and Facebook have created their own 'knowledge graphs' that power semantic searches and enable smarter processing and delivery of data. The use of these knowledge graphs is now the norm rather than the exception" (ISWC 2014)

however, there is no precise definition of knowledge graphs...

Intuitively, **a Knowledge Graph is a knowledge base in the form of graph**

So, what is a Knowledge Graph?

"...major companies such as Google, Yahoo!, Microsoft, and Facebook have created their own 'knowledge graphs' that power semantic searches and enable smarter processing and delivery of data. The use of these knowledge graphs is now the norm rather than the exception" (ISWC 2014)

however, there is no precise definition of knowledge graphs...

Intuitively, **a Knowledge Graph is a knowledge base in the form of graph**

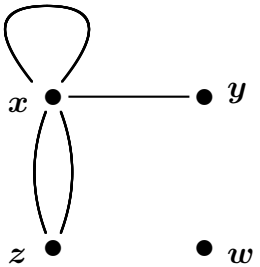
What is a **knowledge base**?

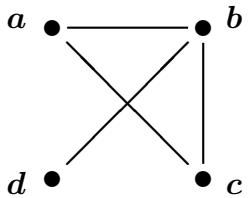
- "A technology to store complex structured and unstructured information used by a computer system. . . represents facts about the world"(Wikipedia)
- "A collection of knowledge expressed using some formal knowledge representation language." (Free Online Dictionary of Computing)
- "A store of information or data that is available to draw on; the underlying set of facts, assumptions, and rules which a computer system has available to solve a problem. (Google Dictionary)

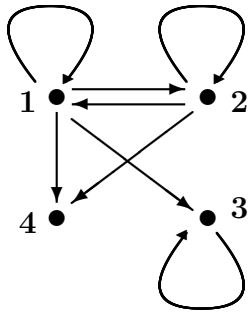
Knowledge bases will be discussed throughout this module

What is a Graph?

Graphs are 'drawings' with dots and (not necessarily straight) lines or arrows:







The dots are called **vertices** (or **nodes**).

The lines or arrows are called **edges**.

Formally, a **graph** is a structure $G = (V, E)$ where V is a non-empty set (of **vertices**) and E a set of (ordered or unordered) pairs of vertices (i.e., **edges**)

Different kinds of graphs

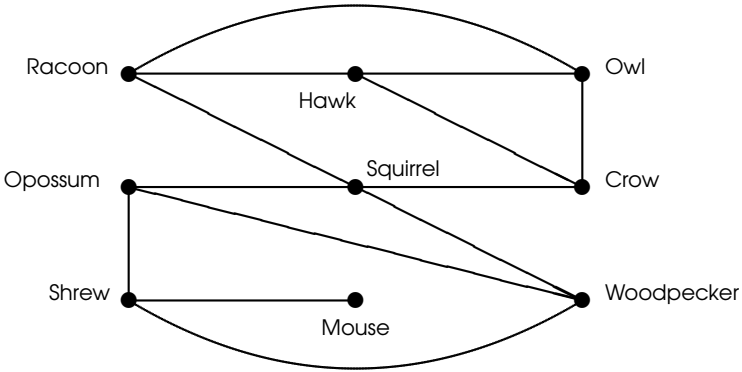
Type	Edges	Multiple edges	Loop edges
(simple) graph	undirected	no	no
multigraph	undirected	yes	yes
directed graph	directed	no	yes
...	...	...	...

Because graphs have applications in a variety of disciplines,
many different terminologies of graph theory have been introduced.

Example 1: Niche overlap graphs in ecology

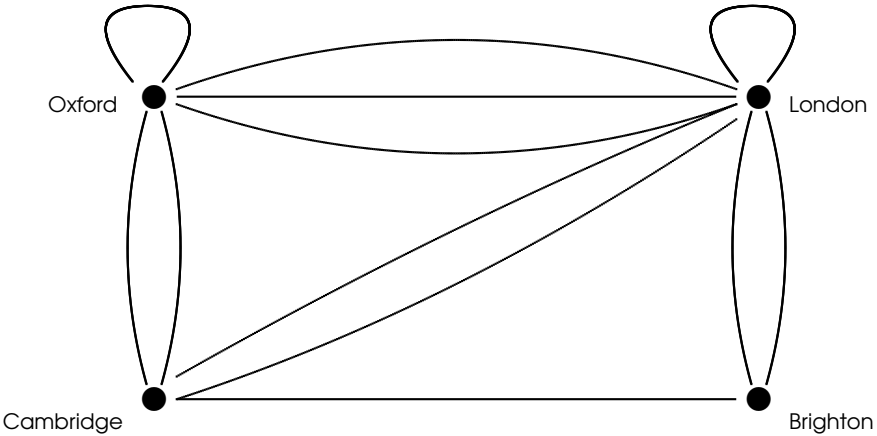
Competitions between species in an ecosystem can be modelled using a **niche overlap graph**:

Each species is represented by a vertex. An edge connects two vertices if the two species represented by these vertices compete (that is, some of the food resources they use are the same).



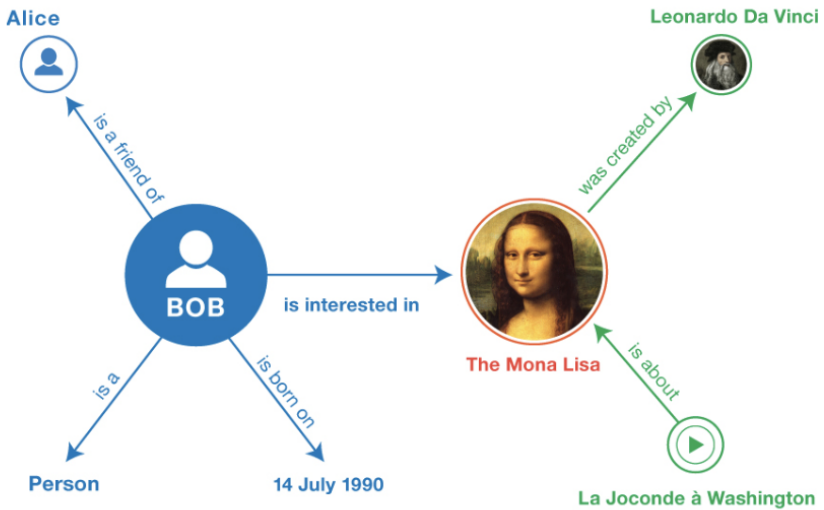
$\rightsquigarrow$ **simple graph** (with labelled vertices)

Example 2: Road networks



$\rightsquigarrow$ **multigraph**

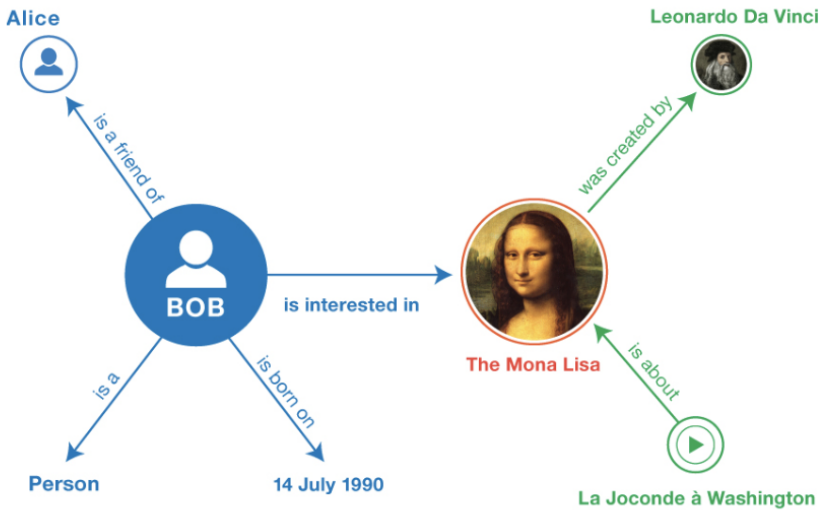
Example 3: ‘Knowledge Graph’



⇝ **directed labelled graph**

What are the labels (in the context of the Web)?

Example 3: ‘Knowledge Graph’



⇝ **directed labelled graph**

What are the labels (in the context of the Web)?

Why ‘graphs’? What about relational databases?

The World Wide Web

⋮

15th century: industrial society, knowledge-based economy

J. Gutenberg developed a moveable type in 1447,
a mechanism to speed the printing of Bibles

⋮

21st century: information society, digital economy

T. Berners-Lee invented the World Wide Web in 1989 at CERN
to provide rapid, electronic access to online technical
reports created by the high-energy physics labs

- social contacts (social networking platforms, blogging, . . .)
- economics (buying, selling, advertising, . . .)
- administration (e-government)
- education (e-learning, . . .)
- etc.

The Semantic Web

TBL's vision of the Web was much more ambitious:

"I have a dream for the Web (in which computers) become capable of analyzing all the data on the Web — the content, links, and transactions between people and computers. A **Semantic Web**, which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The intelligent agents people have touted for ages will finally materialize."

(Berners-Lee, 1999)

The **Semantic Web** is a 'web of data' that facilitates machines to understand the semantics, or meaning, of information on the WWW. It extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other, enabling automated agents to access the Web more intelligently and perform tasks on behalf of users

Berners-Lee is now the director of the World Wide Web Consortium (W3C), which oversees the development of Semantic Web standards. Since 2013, Semantic Web activities have been subsumed by **Web of Data** activities

Understanding the problem with WWW

How can we answer the queries:

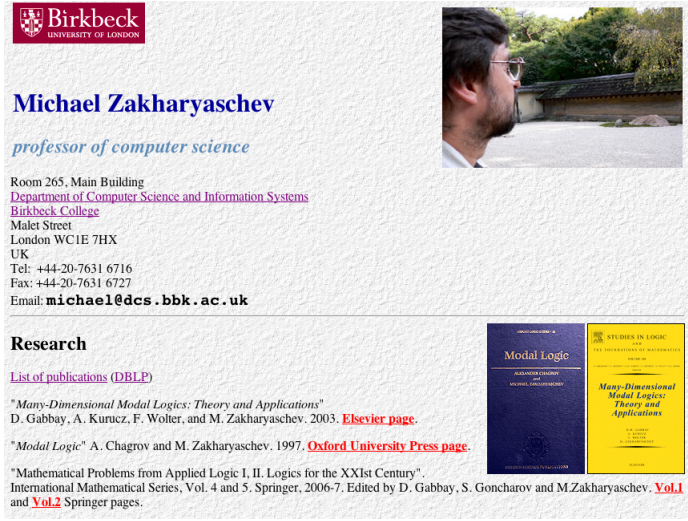
Where does MZ work?
  What is his research area?
Did he publish a book?
  What is his academic position?
. . .

Understanding the problem with WWW

How can we answer the queries:

Where does MZ work?
  What is his research area?
Did he publish a book?
  What is his academic position?
. . .

Google 'Michael Zakharyaschev'



Birkbeck
UNIVERSITY OF LONDON

Michael Zakharyaschev

professor of computer science

Room 265, Main Building
Department of Computer Science and Information Systems
Birkbeck College
Malet Street
London WC1E 7HX
UK
Tel: +44-20-7631 6716
Fax: +44-20-7631 6727
Email: michael@dcs.bbk.ac.uk

Research

List of publications (DBLP)

"*Many-Dimensional Modal Logics: Theory and Applications*"
D. Gabbay, A. Kurucz, F. Wolter, and M. Zakharyaschev. 2003. Elsevier page.

"*Modal Logic*" A. Chagrov and M. Zakharyaschev. 1997. Oxford University Press page.

"Mathematical Problems from Applied Logic I, II. Logics for the XXIst Century".
International Mathematical Series, Vol. 4 and 5. Springer, 2006-7. Edited by D. Gabbay, S. Goncharov and M.Zakharyaschev. Vol.1 and Vol.2 Springer pages.

The Web page contains **enough information** to answer the queries

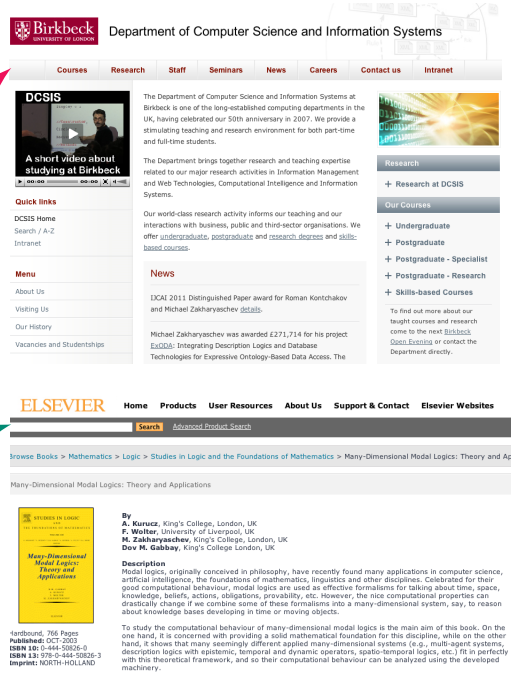
- but this information is **implicit**
- we understand it because we 'know' the **context**
- while machines cannot make sense of it

Task: can we make the data on the Web **explicit** and **machine readable ?**

How to make the data on the Web more accessible?

works at

published by

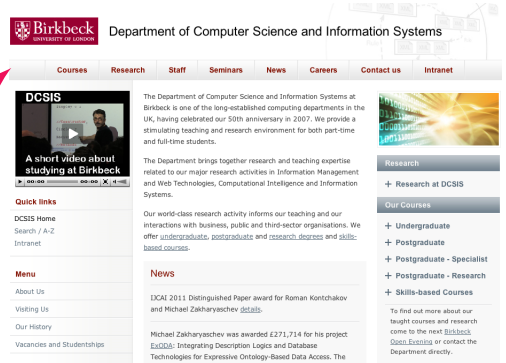


- some **extra information**—**metadata**—must be added to links and data
- this information links data to other data and gives meaning to (characterises) links & data
- this information must be **machine readable**
- this should be done in a **standard way**

How to make the data on the Web more accessible?

works at

published by



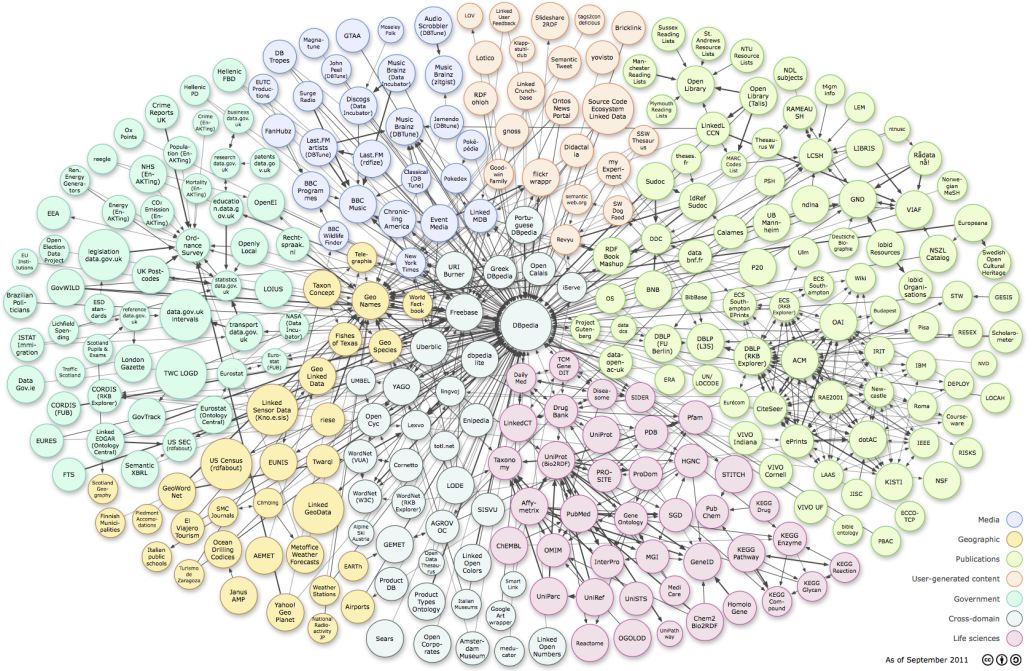
- some **extra information**—**metadata**—must be added to links and data
- this information links data to other data and gives meaning to (characterises) links & data
- this information must be **machine readable**
- this should be done in a **standard way**

Web of Data

'Knowledge Graph'

Linked Data

a method of publishing structured data so that it can be interlinked and become useful through semantic queries. It builds upon standard Web technologies such as HTTP, RDF and URIs, and enables data from different sources to be connected and queried. Linked Data in 2017



Linked Data basic principles

1. Use **URIs** (uniform resource identifiers) to name (identify) things
2. Use **HTTP URIs** so that these things can be looked up (interpreted, 'dereferenced')
3. Provide useful information about what a name identifies when it's looked up, using open standards such as **RDF**, **SPARQL**, etc.
4. Refer to other things using their **HTTP URI-based names** when publishing data on the Web.

- All kinds of conceptual things, they have names now that start with HTTP.
- If I take one of these HTTP names and I look it up, I will get back some data in a standard format which is kind of useful data that somebody might like to know about that thing, about that event.
- When I get back that information it's not just got somebody's height and weight and when they were born, it's got relationships. And when it has relationships, whenever it expresses a relationship then the other thing that it's related to is given one of those names that starts with HTTP.

Another application of KGs: data integration

Bookstore dataset A (relational database)

ID	Author	Title	Publisher	Year
ISBN-0-00-651409-X	id_xyz	The Glass Palace	id_qpr	2000

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

ID	Publisher	City
id_qpr	Harper Collins	London

Bookstore dataset F (Excel sheet)

	A	B	C	D
1	**ID**	**Titre**	**Traducteur**	**Original**
2	ISBN0-20203886682	Le Palais des miroirs	A13	ISBN-0-00-651409-X

...

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

...

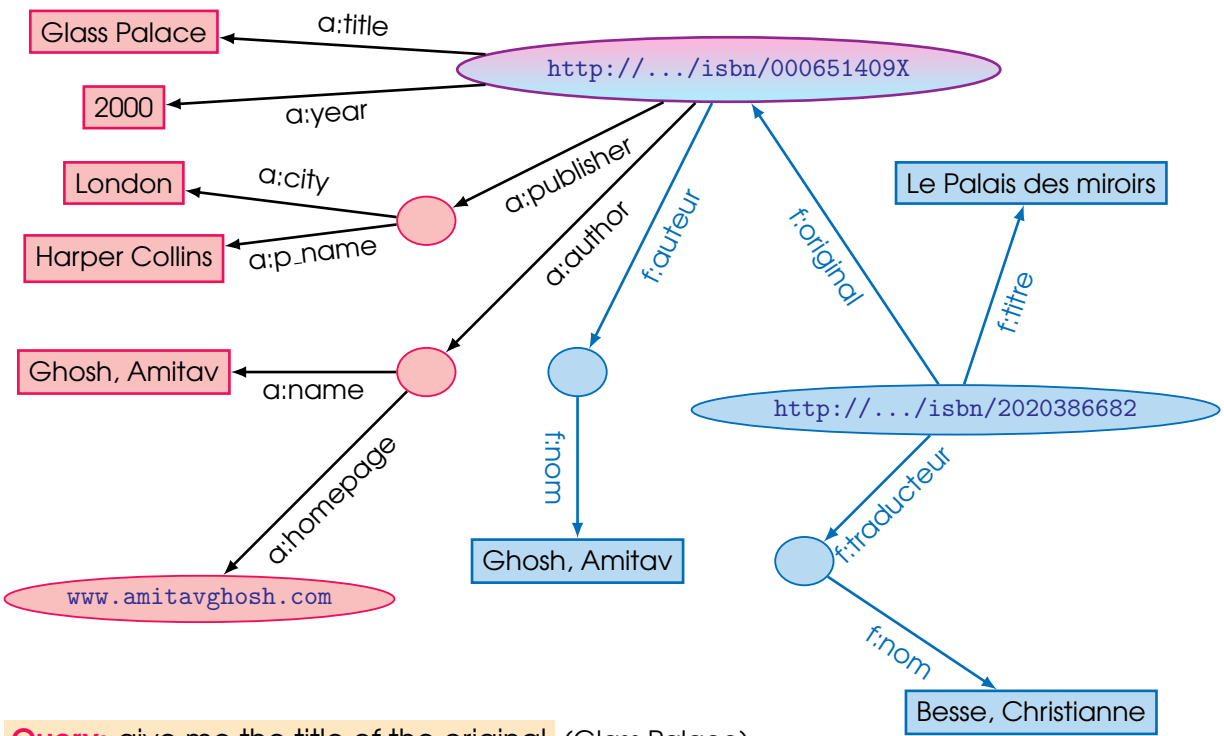
11	**Nom**
12	Ghosh, Amitav
13	Besse, Christianne

Query: what is the title of the original?

(no answer)

Merge in an abstract graph data model

(two identical URIs merged)



Query: give me the title of the original (Glass Palace)

Add more information

The data representation on previous page can be constructed **by the machine**

but the machine doesn't know that a:author and f:auteur should be the same

We can add some extra information to the merged data:

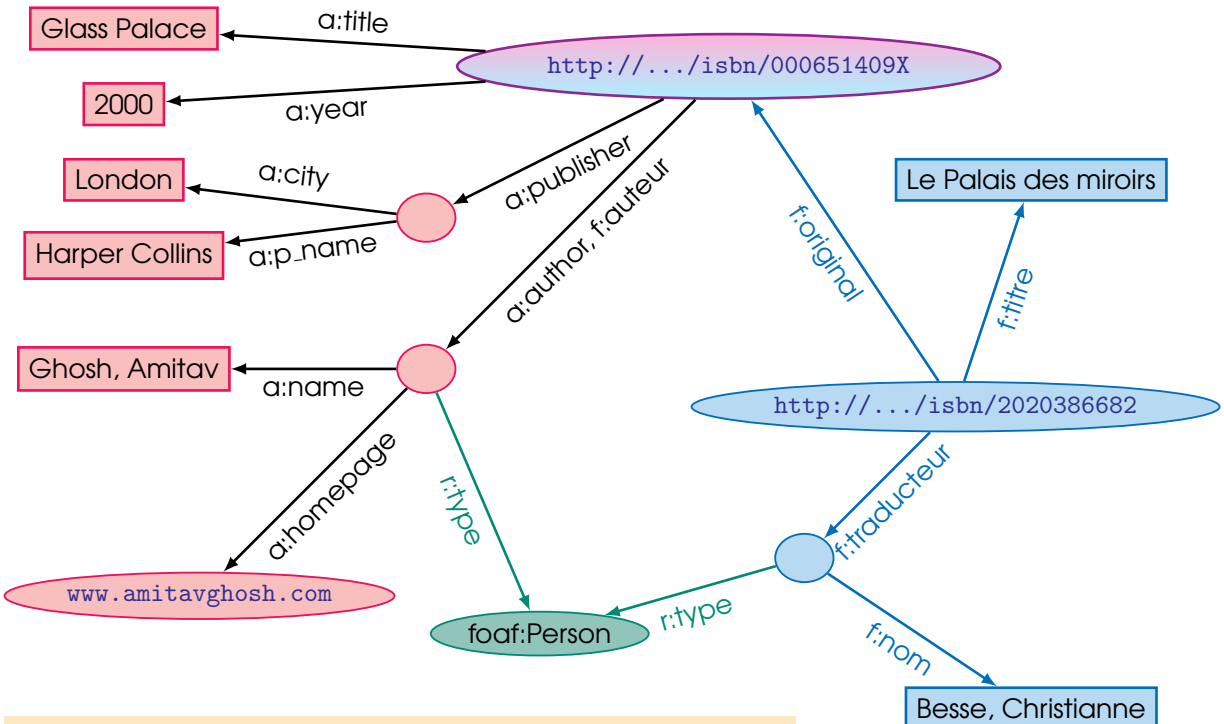
- a:author is equivalent to f:auteur
- both refer to a 'Person' (every a:author is a person)
- the term 'Person' may already be defined by the Web community
- anyway, we may state that
  - a Person is uniquely identified by the name and homepage
  - can be used as a category for certain type of resources

This will provide answers to more queries, e.g.,

Query: give me the home page of the original's auteur

- The dataset can be further combined with other sources such as Wikipedia

Extending merged data



Query: give me the home page of the original's auteur

What did we do?

1. We combined different datasets, which
   - are somewhere on the Web,
   - are of different formats (Mysql, Excel, HTML, etc.),
   - have different names for relations,

   into a "knowledge graph"
2. We could combine the data because some URIs were identical
3. We could add some simple extra information (the 'glue'),
   possibly using common terminologies produced by the community

As a result, new relations could be found and retrieved

It can become even more powerful if we add extra knowledge such as:

- a full classification of various types of library data
- geographical information
- etc.

What are Semantic Technoligies?

Semantic Technologies can be thought of as a collection of standard technologies to realise a Web of Data

The examples above show that we need:

1. formal, machine understandable languages to describe, query, etc. the data and their connections

2. formal `rules' that allow the machines to extract information from the data (classify, query, etc.)

3. corresponding technologies and efficient tools

And apart from that, we need

4. `**ontologies**' in those languages that describe various types of data

In this module, we consider some fundamental aspects of these problems

What is Semantics?

Semantics (Greek *semantikos*, giving signs, significant, symptomatic, from sema, sign) refers to the aspects of meaning that are expressed in a language, code, or other form of representation.

In other words, semantics refers to the meanings assigned to symbols and sets of symbols in a language.

What is Semantics?

Semantics (Greek *semantikos*, giving signs, significant, symptomatic, from sema, sign) refers to the aspects of meaning that are expressed in a language, code, or other form of representation.

In other words, semantics refers to the meanings assigned to symbols and sets of symbols in a language.

Is it hard to explain the meaning of, say, **'a pint of ale'**

- to a human?

What is Semantics?

Semantics (Greek *semantikos*, giving signs, significant, symptomatic, from sema, sign) refers to the aspects of meaning that are expressed in a language, code, or other form of representation.

In other words, semantics refers to the meanings assigned to symbols and sets of symbols in a language.

Is it hard to explain the meaning of, say, **'a pint of ale'**

- to a human?
- to a computer?

What is Semantics?

Semantics (Greek *semantikos*, giving signs, significant, symptomatic, from sema, sign) refers to the aspects of meaning that are expressed in a language, code, or other form of representation.

In other words, semantics refers to the meanings assigned to symbols and sets of symbols in a language.

Is it hard to explain the meaning of, say, **'a pint of ale'**

- to a human?
- to a computer?

Now let's check at

`http://en.wikipedia.org/wiki/Ale`
`http://dbpedia.org/page/Ale`
`https://www.wikidata.org/wiki/Q208385`

analyse the 'explanations'

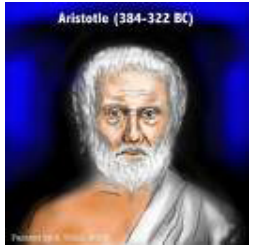
Ontology: origins and history

Ontology in Philosophy

$\mathit{ontologia}$

a philosophical discipline — a branch of philosophy that deals with the nature and the organisation of reality

- Science of Being (Aristotle, Metaphysics, IV, 1)
- Tries to answer the questions:
  - What characterises being?
  - Eventually, what is being?
- How should things be classified?



Ontology in Philosophy

In philosophy, **ontology** is the study of being or existence.

It aims to find out what entities and types of entities exist:

- What exists?
- Is existence a property?
- What is an object?
- Do non-physical (abstract) objects exist?
- How things should be classified?

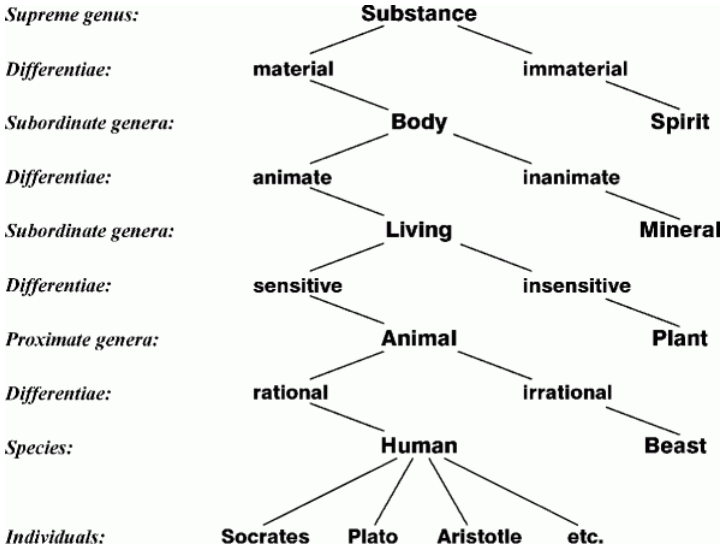
Ontology in Philosophy

In philosophy, **ontology** is the study of being or existence.

It aims to find out what entities and types of entities exist:

- What exists?
- Is existence a property?
- What is an object?
- Do non-physical (abstract) objects exist?
- How things should be classified?

Aristotle's ontology:



Ontology in Computer Science

An ontology is an **engineering artefact**

- It is constituted by a specific **vocabulary** used to describe a certain reality, plus
- a set of **explicit assumptions** regarding the intended meaning of the vocabulary. (Almost always including how concepts should be classified.)

Ontology in Computer Science

An ontology is an **engineering artefact**

- It is constituted by a specific **vocabulary** used to describe a certain reality, plus
- a set of **explicit assumptions** regarding the intended meaning of the vocabulary. (Almost always including how concepts should be classified.)

Thus, an ontology describes a **formal specification** of a certain domain:

- Shared understanding of a domain of interest
- Formal and machine manipulable model of a domain of interest

> *"An explicit specification of a conceptualisation"*
>
> [Tom Gruber 1993]

Schema.org

Schema.org was launched in 2011 by Bing, Google, Yahoo!, Yandex (largest search engines) to **create and support a common set of schemas for structured data markup on web pages**

They propose using the **schema.org** vocabulary along with the Microdata, RDFa, or JSON-LD formats to mark up website content with metadata about itself. Such markup can be recognised by search engine spiders and other parsers, thus gaining access to the **meaning** of the sites.

Inspired by earlier formats such as Microformats, FOAF, OpenCyc.

To test the validity of the data marked up with the schemas and Microdata, such validators as the Google Structured Data Testing Tool, Yandex Microformat validator and Bing Markup Validator can be used.

Some Schema markups such as **Organization** and **Person** are used to influence Google's Knowledge Graph results. `http://schema.org/Person`

How to mark up your content using microdata: `http://schema.org/docs/gs.html`

Google's Knowledge Graph

The **Knowledge Graph** is a knowledge base used by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources. Knowledge Graph display was added to Google's search engine in 2012.

It uses a graph database to provide structured and detailed information about the topic in addition to a list of links to other sites. The goal is that users would be able to use this information to resolve their query without having to navigate to other sites and assemble the information themselves. The short summary provided in the knowledge graph is often used as a spoken answer in Google Assistant searches.

According to some news websites, the implementation of Google's Knowledge Graph has played a role in the page view decline of various language versions of Wikipedia. As of the end of 2016, knowledge graph holds over 70 billion facts.

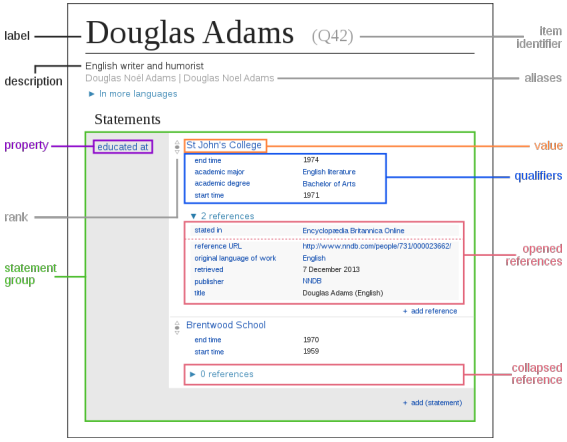
https://www.google.com/intl/bn/insidesearch/features/search/knowledge.html

Wikidata

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines. Wikidata acts as central storage for the structured data of its Wikimedia sister projects including

Wikipedia, Wikivoyage, Wikisource, and others.

Wikidata is a document-oriented database, focused on items. Each item represents a topic and is identified by a unique ID. Information is added to items by creating statements. Statements take the form of key-value pairs.

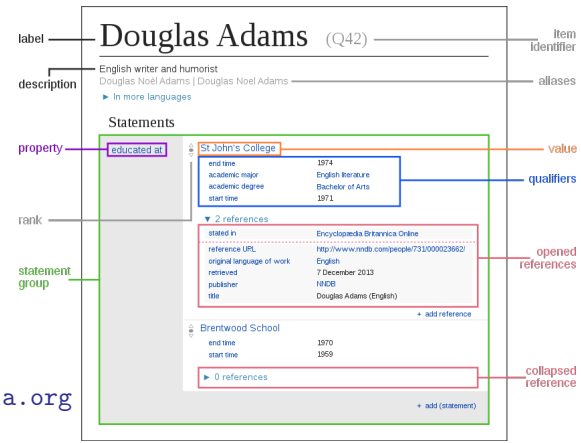


Wikidata

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines. Wikidata acts as central storage for the structured data of its Wikimedia sister projects including

Wikipedia, Wikivoyage, Wikisource, and others.

Wikidata is a document-oriented database, focused on items. Each item represents a topic and is identified by a unique ID. Information is added to items by creating statements. Statements take the form of key-value pairs.



also `http://wiki.dbpedia.org`

Ontologies in sciences

- Bioinformatics
  - The Gene Ontology, The Protein Ontology MGED, etc.
- Medicine
  - The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) Ontology: a Core terminology of over 364,000 health care concepts; more than 984,000 descriptions; $\approx$ 1.45 million semantic relationships.

Pericardium **is-a** *Tissue* **and** *containedIn* . *Heart*
Pericarditis **is-a** *Inflammation* **and** *hasLocation* . *Pericardium*
Inflammation **is-a** *Disease* **and** *actsOn* . *Tissue*
Disease **and** *hasLocation* . *containedIn* . *Heart* **is-a** *HeartDisease* **and** *NeedsTreatment*

- Linguistics
- Database integration
- User interface design
- Fractal Indexing
- ...

Semantic Technologies at the BBC

BBC Online

Launched in the mid 1990s, the BBC website was focused on supporting

- broadcast brands such as Top Gear as well as
- domain-specific sites: news, food, gardening, etc.

BBC Web-based service is one of the most visited websites and the world's largest news website. As of 2007, it contained over two million pages

BBC Online

Launched in the mid 1990s, the BBC website was focused on supporting

- broadcast brands such as Top Gear as well as
- domain-specific sites: news, food, gardening, etc.

BBC Web-based service is one of the most visited websites and the world's largest news website. As of 2007, it contained over two million pages

Focus has been on **separate, standalone HTML microsites** that are **not linked** together and to other data sources on the Web

→ difficult to find everything BBC has published about any given object

cannot navigate from a page about a musician to a page with all the programmes that have played that artist, to their biography, etc.

Creating a website for the Football World Cup 2010

32 teams, 8 groups, 776 players

too many pages to create, too few journalists to create & manage content

Solution use Semantic Technologies:

- ontology describes the interrelation between facts of the World Cup
- all such metadata stored as RDF triples

Example: 'Frank Lampard' is part of 'England Squad'

'England Squad' competes in 'Group C' of the 'FIFA World Cup 2010'

> *"The underlying publishing framework does not author content directly; rather it publishes data about the content — metadata. The published metadata describes the world cup content at a fairly low-level of granularity, providing rich content relationships and semantic navigation. By querying this published metadata we are able to create dynamic page aggregations for teams, groups and players."*

Jem Rayfield, Senior Technical Architect, BBC News and Knowledge

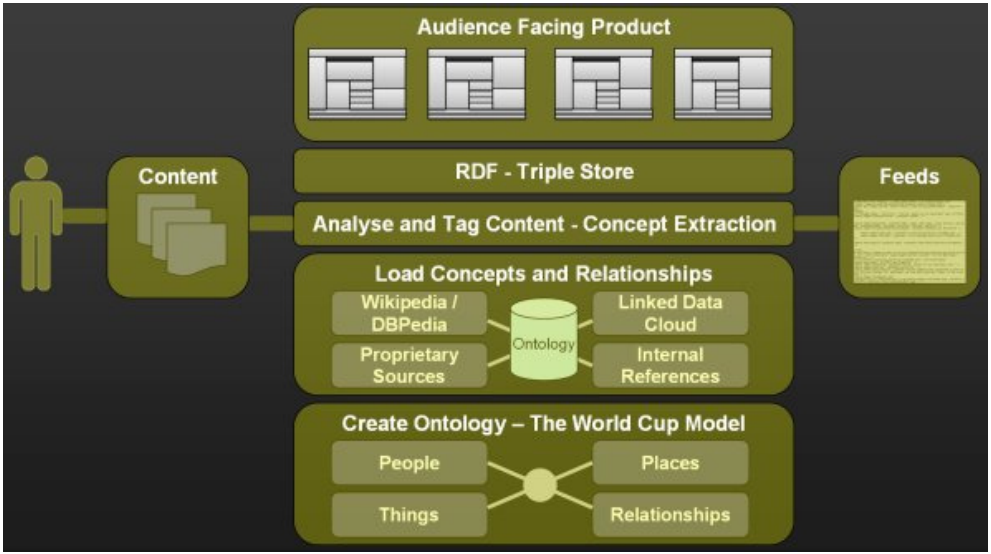
http://www.bbc.co.uk/blogs/bbcinternet/2010/07/bbc_world_cup_2010_dynamic_sem.html

The BBC website for the Football World Cup 2010

- Inference for enrichment of the data and SPARQL for queries
- In addition, the ontology contains parts written by journalists: stories, blogs, profiles, images, videos and strategies
- Journalistic articles are tagged automatically (NLP techniques) and manually
- Stats and scores from other sources are imported from XML and mapped to ontological concepts
- Web pages are created automatically and contain relevant references
- Use of the technique also for the 2012 Olympic Games in London

The BBC Football World Cup 2010

The underlying architecture



- Information is dynamically aggregated from external, publicly available data
- All data available as Linked Open Data
- Data access via simple HTTP request
- Data is always up-to-date without manual interaction

Data access in industry

(from Norwegian Petroleum Directorate's FactPages)

show me the wellbores completed before 2008 where Equinor as a drilling operator sampled less than 10 meters of cores

Data access in industry

(from Norwegian Petroleum Directorate's FactPages)

show me the wellbores completed before 2008 where Equinor as a drilling operator sampled less than 10 meters of cores

5 days later:

```
SELECT DISTINCT cores.wlbName, cores.lenghtM, wellbore.wlbDrillingOperator, wellbore.wlbCompletionYear
FROM
    ( (SELECT wlbName, wlbNpdidWellbore, (wlbTotalCoreLength * 0.3048) AS lenghtM
       FROM wellbore_core
       WHERE wlbCoreIntervalUom = '(ft )'  )
      UNION
      (SELECT wlbName, wlbNpdidWellbore, wlbTotalCoreLength AS lenghtM
       FROM wellbore_core
       WHERE wlbCoreIntervalUom = '(m )'  )
    ) as cores,
    ( (SELECT wlbNpdidWellbore, wlbDrillingOperator, wlbCompletionYear
       FROM wellbore_development_all
      UNION
      (SELECT wlbNpdidWellbore, wlbDrillingOperator, wlbCompletionYear
       FROM wellbore_exploration_all  )
      UNION
      (SELECT wlbNpdidWellbore, wlbDrillingOperator, wlbCompletionYear
       FROM wellbore_shallow_all  )
    ) as wellbore
WHERE wellbore.wlbNpdidWellbore = cores.wlbNpdidWellbore
```

...

Data access in industry

(from Norwegian Petroleum Directorate's FactPages)

show me the wellbores completed before 2008 where Equinor as a drilling operator sampled less than 10 meters of cores

5 days later:

```
SELECT DISTINCT cores.wlbName, cores.lenghtM, wellbore.wlbDrillingOperator, wellbore.wlbCompletionYear
FROM
    ( (SELECT wlbName, wlbNpdidWellbore, (wlbTotalCoreLength * 0.3048) AS lenghtM
       FROM wellbore_core
       WHERE wlbCoreIntervalUom = '(ft )' )
      UNION
      (SELECT wlbName, wlbNpdidWellbore, wlbTotalCoreLength AS lenghtM
       FROM wellbore_core
       WHERE wlbCoreIntervalUom = '(m )' )
    ) as cores,
    ( (SELECT wlbNpdidWellbore, wlbDrillingOperator, wlbCompletionYear
       FROM wellbore_development_all
      UNION
      (SELECT wlbNpdidWellbore, wlbDrillingOperator, wlbCompletionYear
       FROM wellbore_exploration_all )
      UNION
      (SELECT wlbNpdidWellbore, wlbDrillingOperator, wlbCompletionYear
       FROM wellbore_shallow_all )
    ) as wellbore
WHERE wellbore.wlbNpdidWellbore = cores.wlbNpdidWellbore
    ...
```

at Equinor (former Statoil):

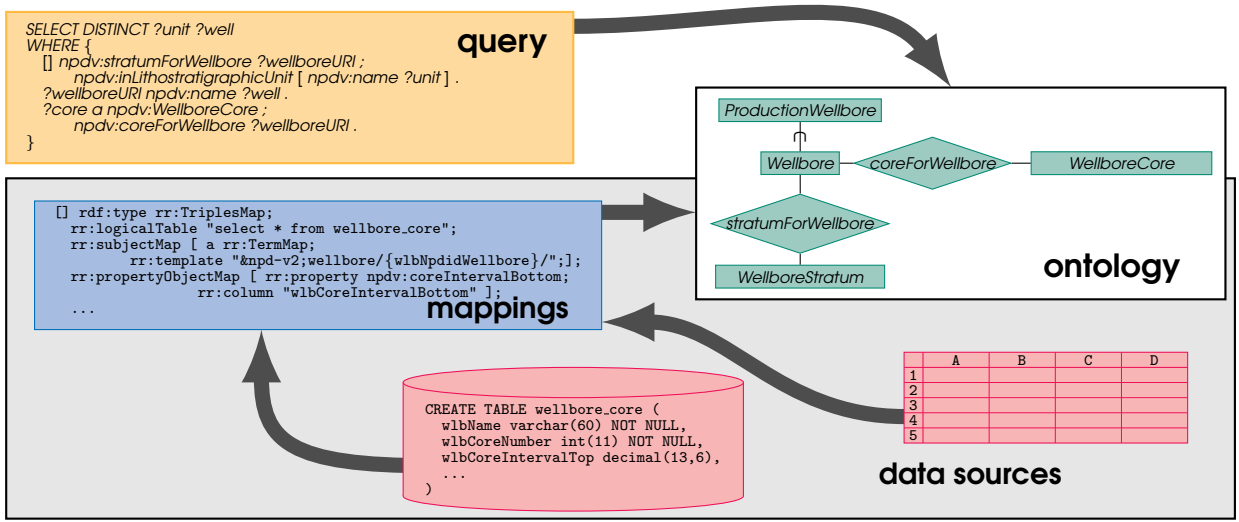
1,000 TB of relational data

2,000 tables

different schemas

30–70% of time on data gathering

Ontology-based data access (OBDA)



Ontology

- gives a high-level conceptual view of the data
- provides a convenient & natural vocabulary for user queries
- enriches incomplete data with background knowledge