

1 Introduction

Course Organization

<https://cw.fel.cvut.cz/wiki/courses/b4m36osw>

1.1 Sharing meaning of data

1.1.1 Some examples of misunderstanding

One event or two events?

DID YOU KNOW



Awaken the mind.

Just months before 9/11, the World Trade Center's lease was privatized and sold to Larry Silverstein.

Silverstein took out an insurance plan that 'fortuitously' covered terrorism.

After 9/11, Silverstein took the insurance company to court, claiming he should be paid double because there were 2 attacks.

Silverstein won, and was awarded \$4,550,000,000.

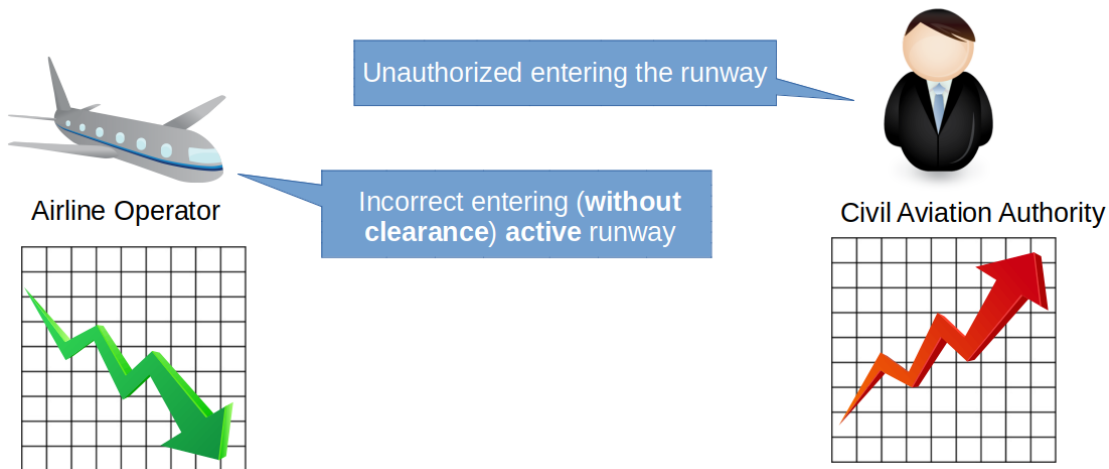
What is an event ? How many events occurred at 9/11 – One or Two ?

Knowledge Management

9/11 ... matter of billions of USD

source:<https://www.metabunk.org/larry-silversteins-9-11-insurance.t2375>

What is the trend of Runway Incursion incidents at an airline operator ?



1.1.2 What is a dataset about?

What is inside a dataset?



See OpenData portal of Prague OpenData portal of Prague

What is a building?

Building is a construction

- both above and below ground
- spatially compact
- with walls and roof
- with heating

Act 406/2000 Coll., on Energy Management

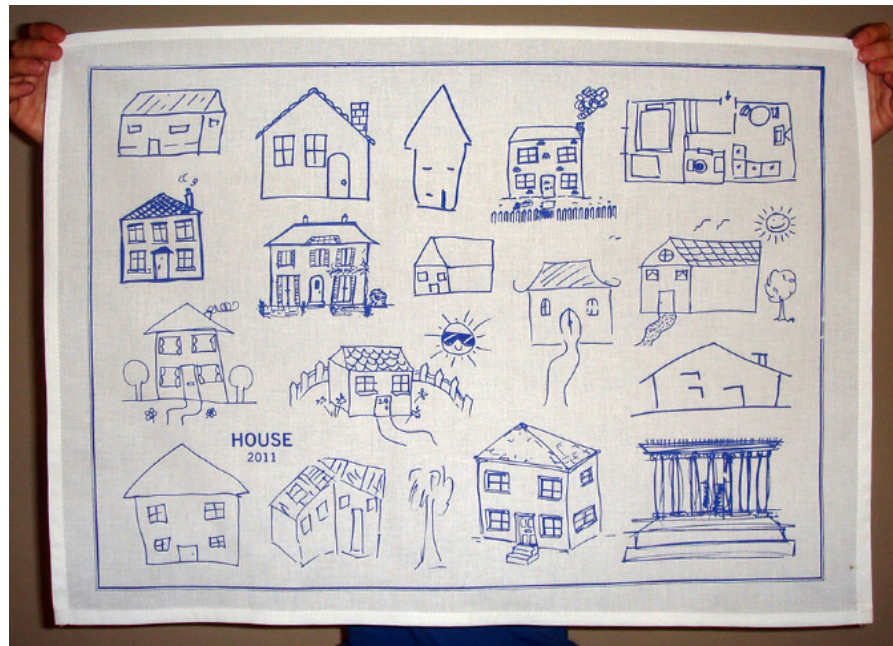
Building is a construction

- above ground
- with solid foundations
- spatially compact
- with walls and roof

Act 256/2013 Coll, Cadastral Law

1 Introduction

But things are even more complicated ...



What is a building?



1. ... is a **construction** which is **heated**.

2. ... is a **construction** to provide **protection** to their users or internal equipment and is typically **closed** and has a **permanent position**.

ČSN EN 15643-5 -Sustainability of construction works

Building

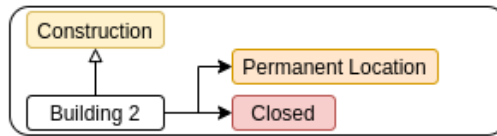
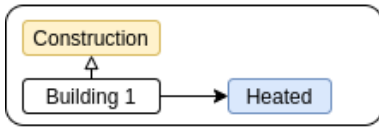
3. ... is a **construction above ground** which is **spatially-compact** and **closed by walls and roof**.

Act 256/2013 Coll., Cadastral Law

4. ... is a **construction above and below ground** which is **spatially-compact** and **closed by walls and roof** and is **heated or cooled**.

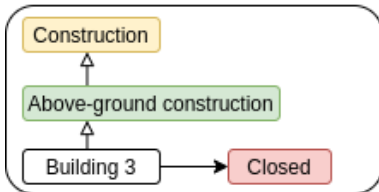
Act 406/2000 Coll., on Energy Management

What is a building?

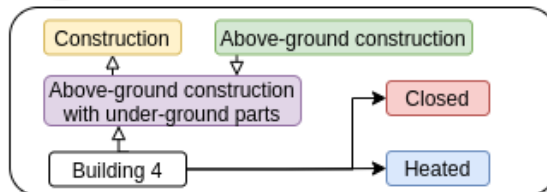


ČSN EN 15643-5 -Sustainability of construction works

Building

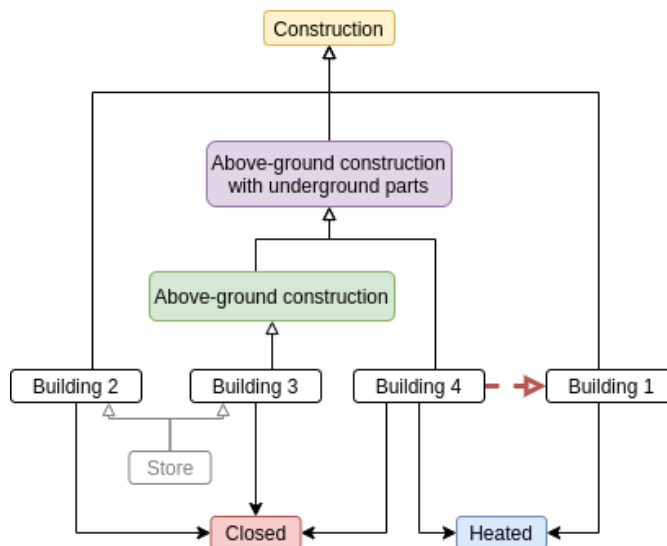


Act 256/2013 Coll., Cadastral Law



Act 406/2000 Coll., on Energy Management

New knowledge can be inferred



1.1.3 From conceptual models to ontologies

Ontological Conceptual Modeling

- a way to **capture** and **explain** meaning.
- the language must be understandable to non-experts (UML max)
- the language must be computable – we want to use the models to infer new knowledge or validate data

1 Introduction

About ontologies

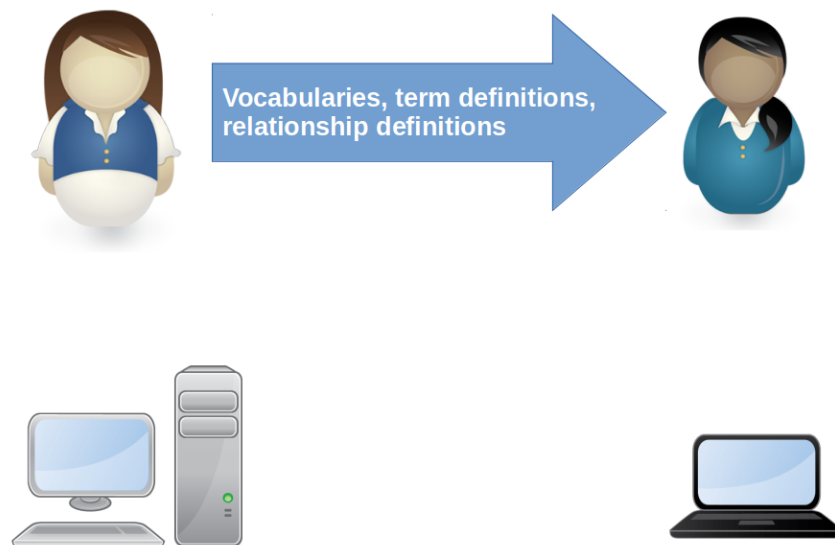
Ontologies

are **formal specifications of conceptualization**.

Ontologies help to stabilize the knowledge, to share meaning both among computers and among people. Use-cases include

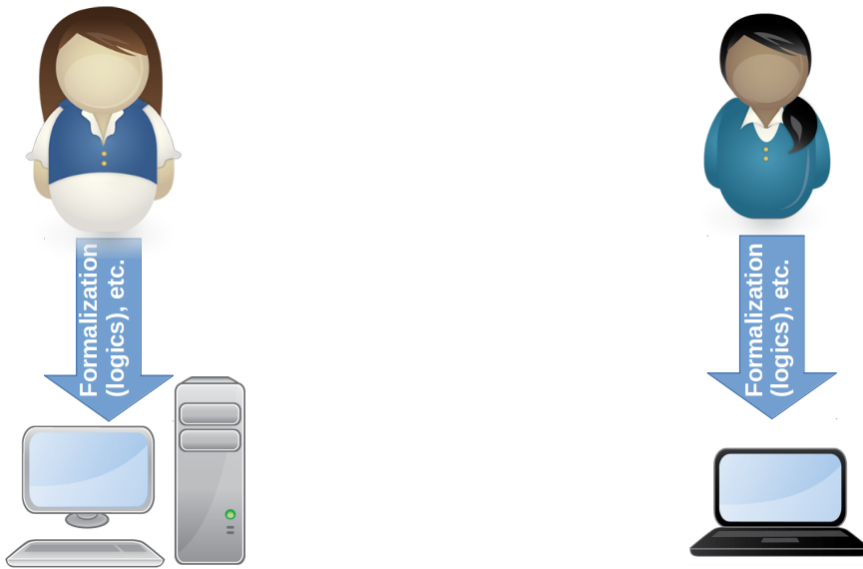
- Data Integration
- Semantic Web
- Open (Linked) Data

First, People Need to Understand Each Other

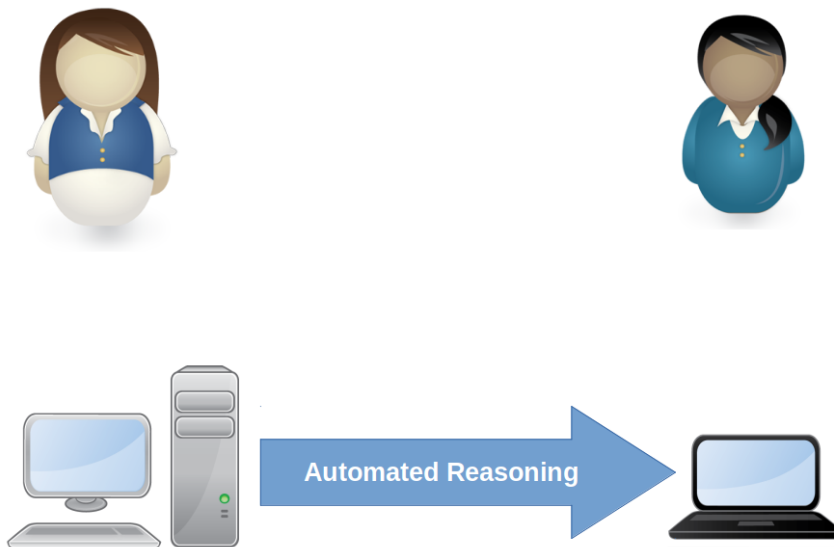


Second, People Need to Explain Things to Computers

1.1 Sharing meaning of data



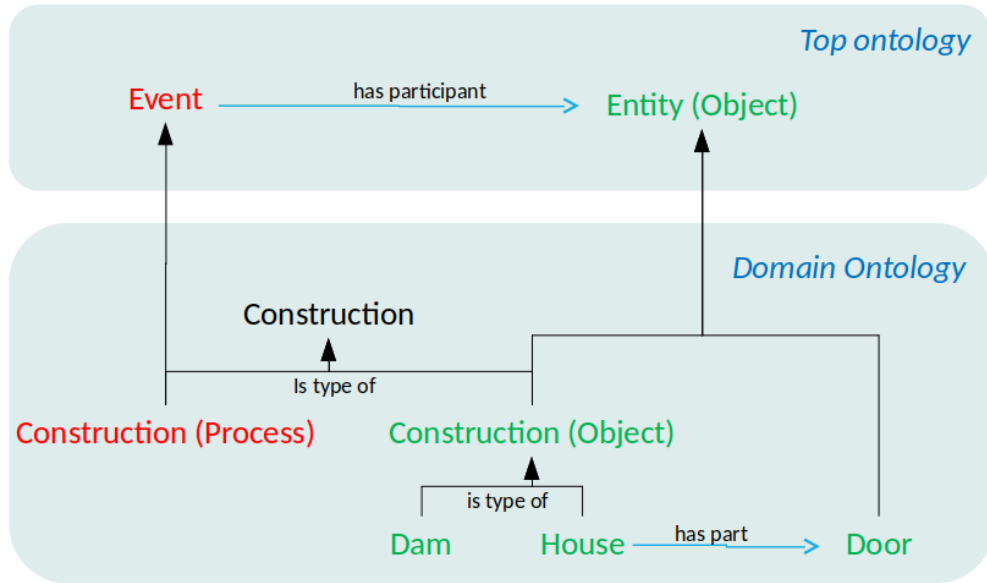
Third, Computers Can Understand One Another



Ontology

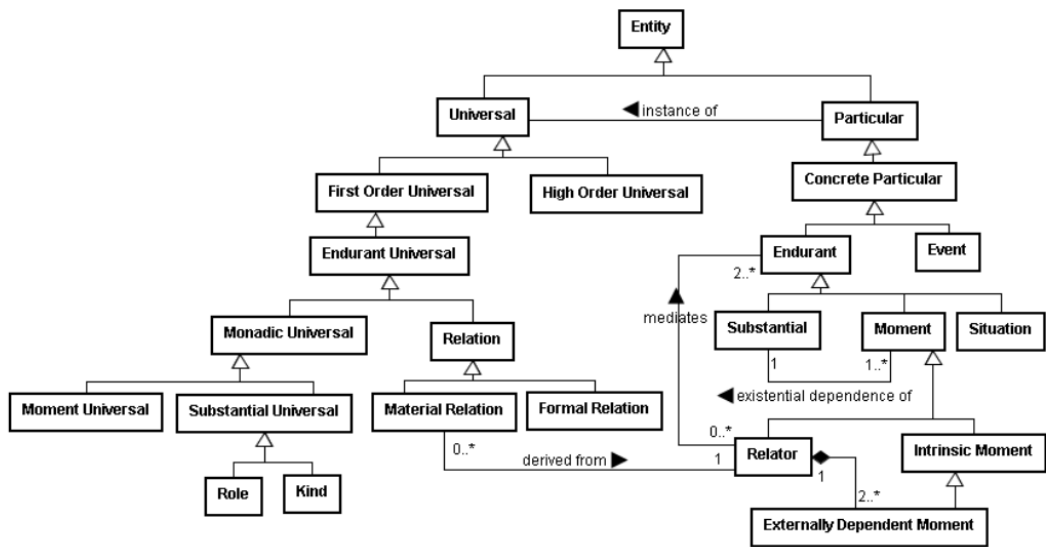
Explicit Conceptualization of Shared Meaning

1 Introduction



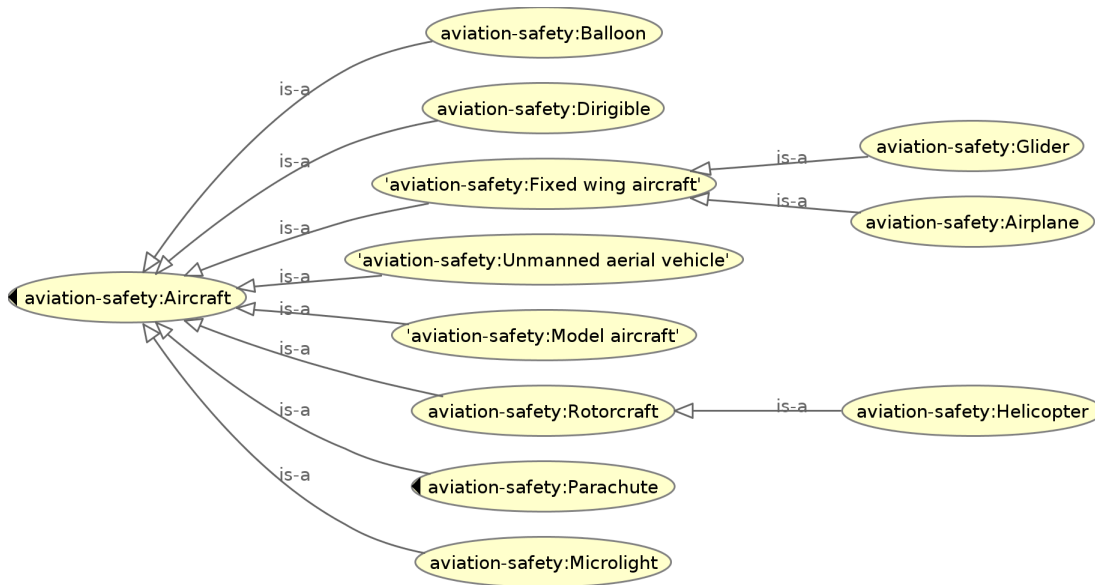
Example Top-Level Ontology

Small part of Unified Foundational Ontology (UFO)



Example Ontology Hierarchy

Each helicopter is also an aircraft.



Ontologies \neq Taxonomies

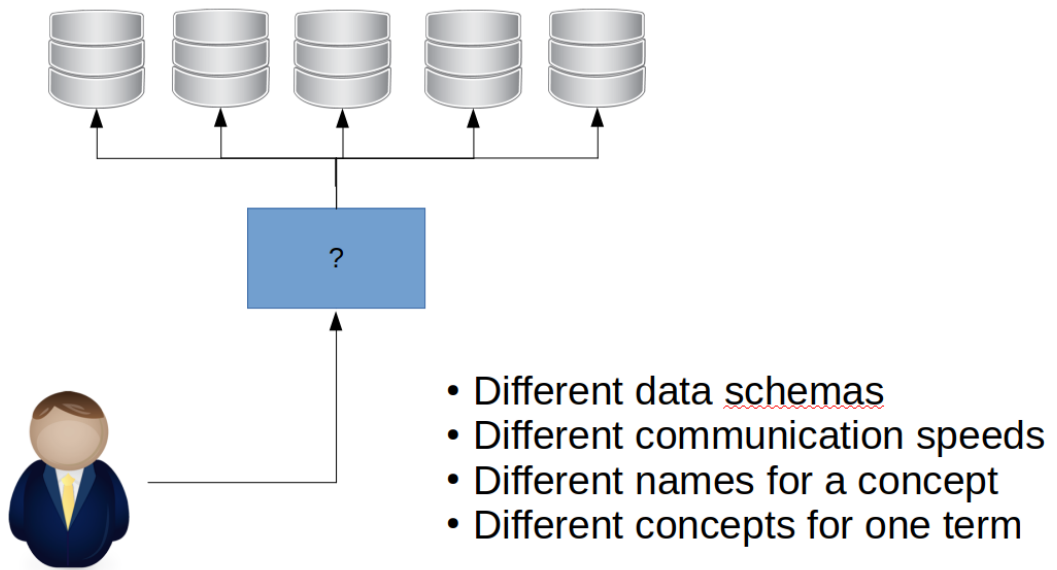
Taxonomies = just a single type of relationship.

Construction	→ broad meaning (object, construction site, process)
Dam	
House	→ broad meaning (dwelling, construction)
Door	→ specific meaning (not type of house, but its part)

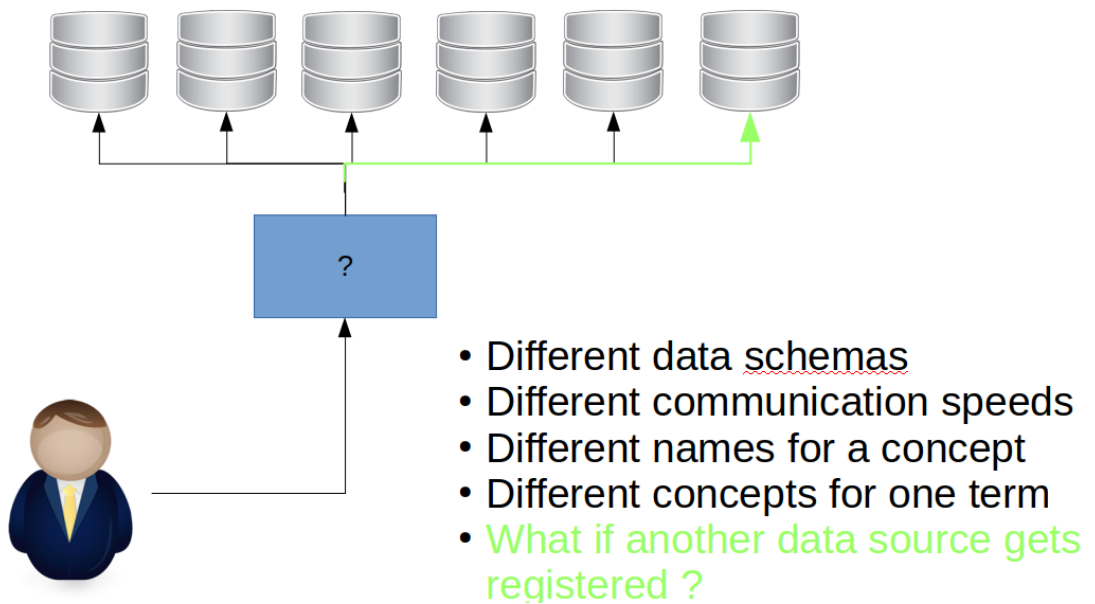
1.1.4 Ontologies for data integration

Data Integration Scenario

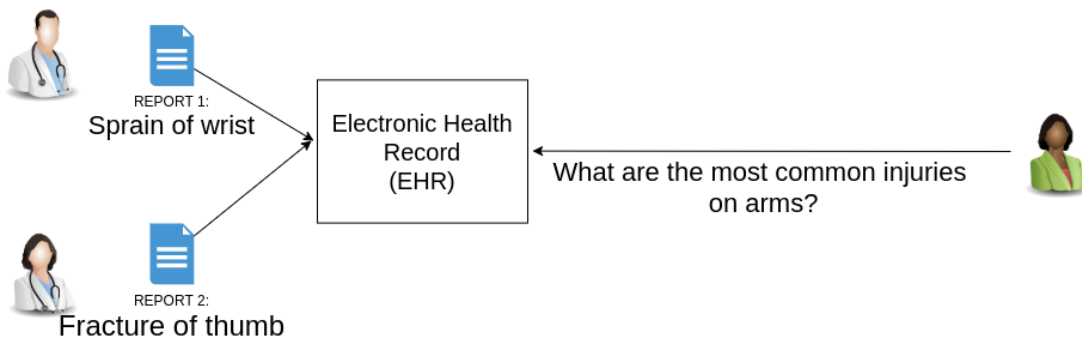
1 Introduction



Data Integration Scenario



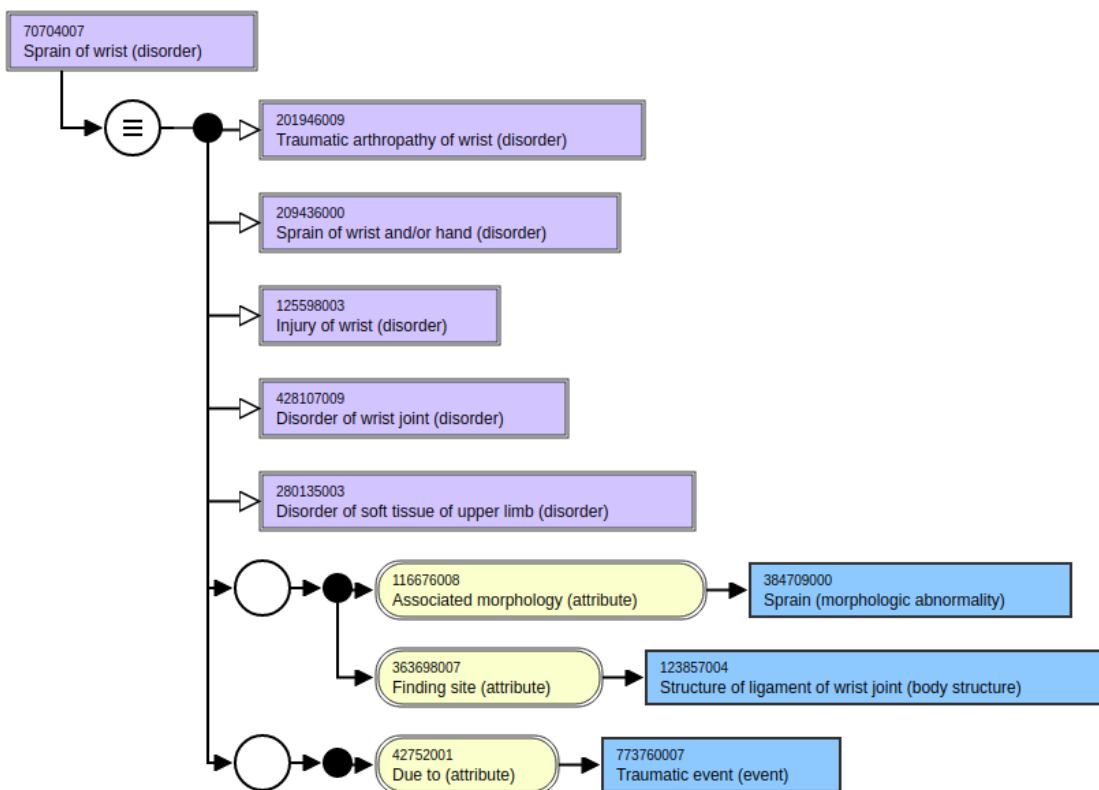
Use-case – HealthCare Data Integration



SNOMED-CT

Systematized Nomenclature of Medicine - Clinical Terms

- ~ 300k clinical concepts
- international standard – adopted e.g. in UK, USA, Australia
- uses ontology reasoning to classify/query the concepts



SNOMED-CT

Systematized Nomenclature of Medicine - Clinical Terms

1 Introduction

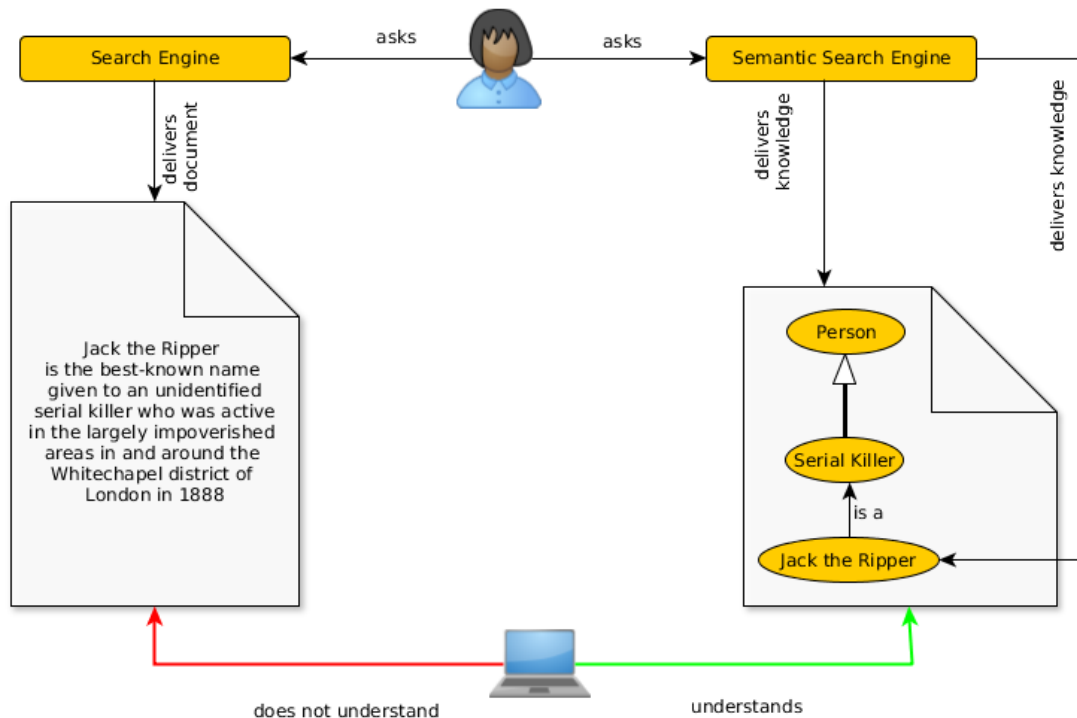
<https://browser.ihtsdotools.org/?perspective=full&conceptId1=70704007&edition=MAIN/2020-07-31&release=&languages=en>

1.2 Semantic Web

Current Web vs. Semantic Web

- SoA – semistructured HTML or XML data. There is vast amount of search engines like Google, Yahoo, MSN, etc. Many of them are invaluable, but as the engines use just keywords and/or some natural language preprocessing methods, the search results contain lots of irrelevant results that need to be processed manually.
- How to make web search more efficient ?
 - more expressive power for web designers to capture complexities – SW languages (RDF(S), OWL),
 - more efficient search engines to handle SW languages – new inference techniques for these languages,
 - better search engines interfaces – more expressive query languages
- **the amount of (unstructured) data is steadily growing**

Semantic search



Ontologies and Semantic Web

ontology has many definitions, but let's consider it a **formal representation of a complex domain knowledge that is shared with others to ensure intelligent system interoperability,**

semantic web is *an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.* (cit. Semantic Web. Tim Berners-Lee, James Hendler and Ora Lassila, Scientific American, 2001)

Idea of Semantic Web

- W3C web page - <http://www.w3.org/2001/sw>
- The data format will be either RDF(S) or OWL,
- Reasoners for RDF(S) can be used for partial derivation in OWL,
- Reasoners for OWL can be used for derivation in RDF(S)

Unique Data Identification – URIs

Semantic web speaks about resources.

URI is a unique identifier for addressing web resources in the form

```
<scheme name> : <hier. part> [ ? <query> ] [ # <fragment> ]
```

. HTTP scheme is used typically.

URN a URI with *scheme name* equal to 'urn'; used e.g. in SWRL atom identification,

URL a URI that can be resolved to a content using the protocol (e.g. HTTP),

IRI generalization of URIs allowing non-ascii characters. IRI is the standard identifier for OWL.

Open World Assumption

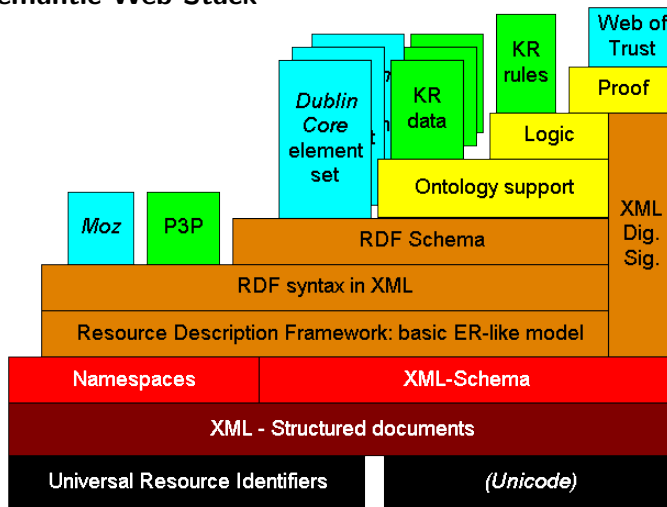
The semantic web inference must take into account that we handle *incomplete knowledge*.

Description

Open world (OWA): Everything that cannot be proven is unknown, Closed world (CWA): Everything that cannot be proven is false.

Statement : "John is a Man." Query: "Is Jack a Man ?" OWA Answer: "I don't know." CWA Answer: "No."

Semantic Web Stack



Taken from <http://www.w3.org/2000/Talks/0906-xmlweb-tbl/slide9-0.html>, by Tim Berners Lee.

1.2.1 Linked Data

How to publish data related to other ?

Based on semantic web principles, Linked Data provide means to efficiently connect data created by different publishers.

- Web of Documents – WWW
 - webpage – readable by human
 - identifiers – IRI
 - transfer protocol – HTTP
 - unified language – HTML
- Web of Data – Linked Data
 - webpage – readable by machine
 - identifiers – IRI
 - transfer protocol – HTTP
 - unified language – RDF

Linked Data [Heath2011] is a method for publishing structured and interlinked data on the web, building up on URIs, HTTP and RDF technologies.

Linked Data Principles

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.

3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things.

(Tim Berners-Lee, 2009 – <http://www.w3.org/DesignIssues/LinkedData.html>)

URIs satisfying the third point are **dereferencable**.

Document vs. its Content

When designing a URI scheme it is necessary to ensure proper distinction between a **document** and its **content**

Example

```
@prefix people: <http://example.com/people/>
people:John people:likes people:Mary
```

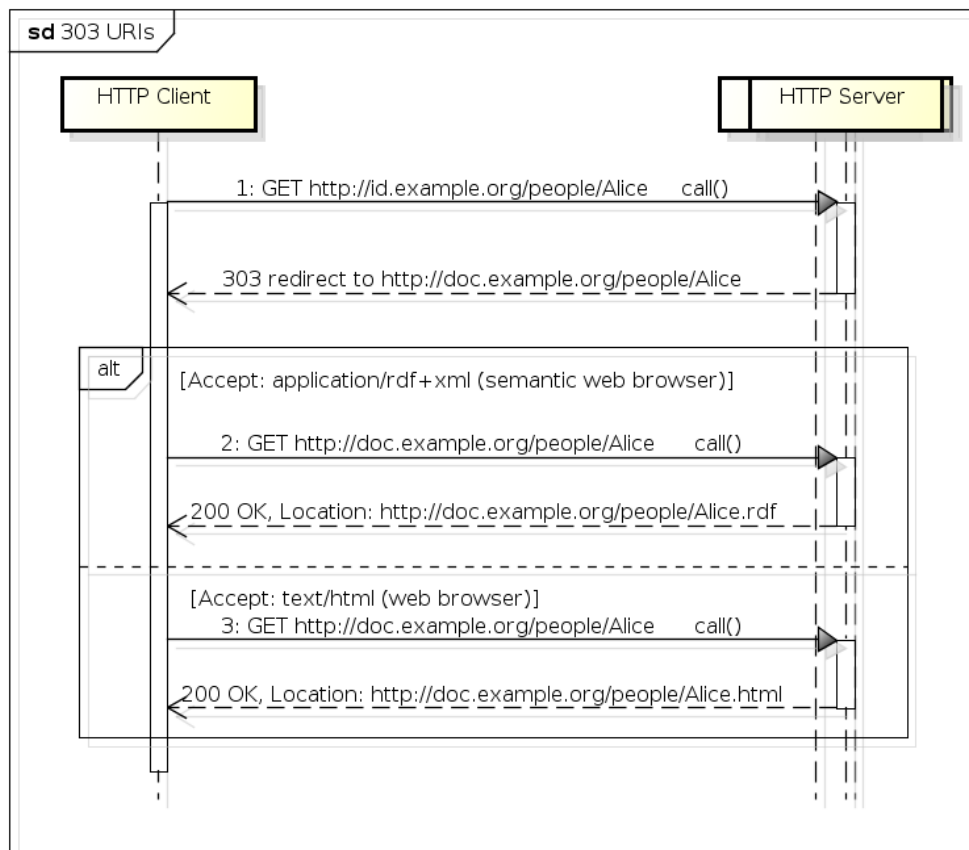
Is <http://example.com/people/Mary> a web document or a resource ? (Consider semantic consequences of each option).

This is handled by two strategies – 303 URIs and Hash URIs, each being suitable for different scenarios.

303 URIs

- 303 URIs are of the form <http://id.example.org/people/Alice>
- HTTP server sends 303 redirect to the corresponding **document** of the requested **resource**.
- HTTP client makes another request, based on Accept headers, the RDF/HTML version is delivered.

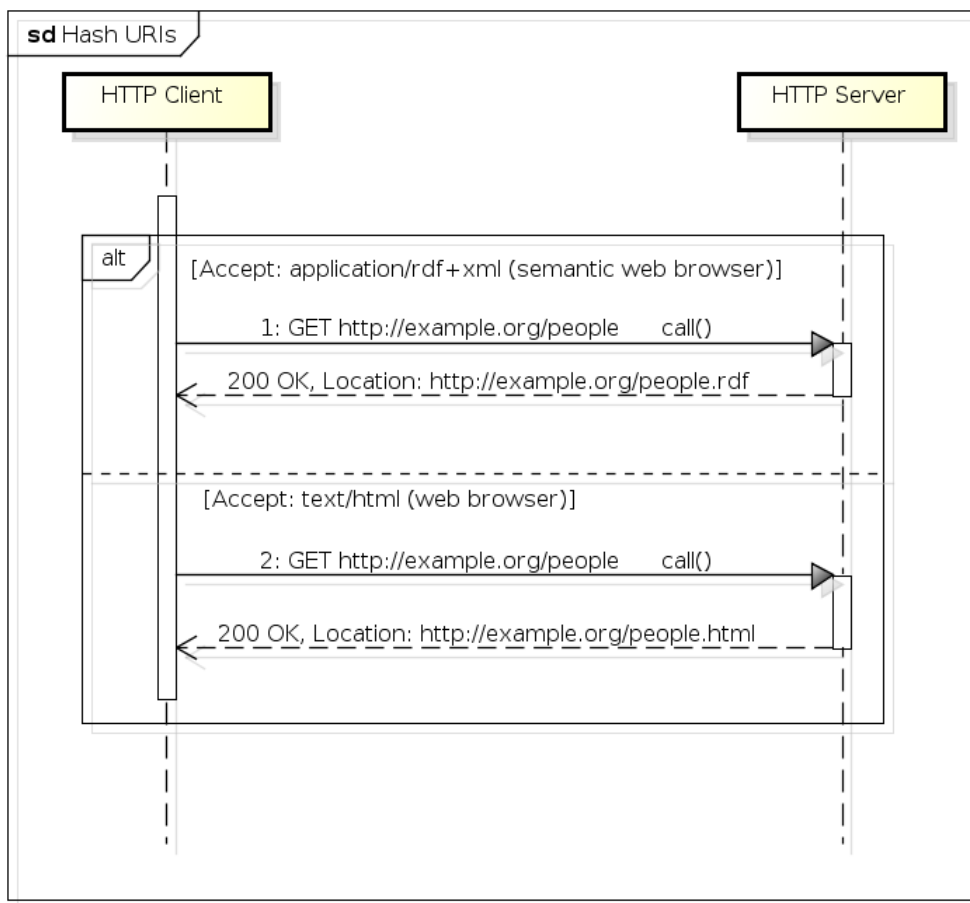
1 Introduction



powered by Astah

Hash URIs

- Hash URIs are of the form `http://example.org/people#Alice`
- HTTP server sends the whole **document** of either RDF or HTML type based on Accept headers.
- Within the document, the HTTP client gets the particular entity after the hash symbol.



powered by Astah

303 URIs vs. Hash URIs

Hash URIs are suitable for small datasets that will hardly grow up,

303 URIs are suitable for large datasets for the sake of good performance.

Reason

The fragment part of an URL (after #) is evaluated on the HTTP client (not the HTTP server), so the HTTP client must fetch all data first and then filter them for the subsequent use locally.

Linked data presentation

LodView (2023, <https://github.com/LodLive/LodView>) a publication tool for Linked Data

1 Introduction

Marmotta (2018, <https://github.com/apache/marmotta>) is a platform for publishing Linked Data (contributed from Linked Media Framework),

Callimachus (2017, <https://github.com/3-Round-Stones/callimachus>) is an application server for linked data applications. To be explored in the tutorials,

D2R (2015, <https://github.com/d2rq/d2rq>) is a platform for publishing relational database data in the form of Linked Data.

Pubby (2014, <https://github.com/cygri/pubby>) is a simple Linked Data publication server connectable to SPARQL endpoints,

1.2.2 Use-case: Open Data

CKAN and DataHub

CKAN (<http://ckan.org/>) is an open-source data portal for publishing, sharing and search of datasets.

It is prominently hosted at <http://datahub.io>. Datasets on DataHub can be sub-

The screenshot shows the DataHub website interface. At the top, there is a navigation bar with the DataHub logo and the tagline 'The easy way to get, use and share data'. The main navigation includes 'Datasets', 'Organizations', 'About', 'Blog', and 'Help'. A search bar is located on the right side of the navigation bar.

The main content area is titled '/ Datasets'. On the left side, there are two filters: 'Organizations' and 'Tags'. The 'Organizations' filter lists various organizations with their respective dataset counts, such as 'Global (5)', 'Linking Open Data C... (2)', and 'VU University Amste... (1)'. The 'Tags' filter lists various tags with their respective dataset counts, such as 'lod (6)', 'culturalheritage (6)', and 'publications (4)'. Both filters have a 'Clear All' link.

The main search results area shows a search for 'cultural heritage'. A search bar contains the text 'cultural heritage'. Below the search bar, there is a button labeled 'Add Dataset'. The search results are displayed as a list of datasets. The first result is '14 datasets found for "cultural heritage"', with an 'Order by: Relevance' dropdown menu. Below this, there are three dataset entries:

- Swedish Open Cultural Heritage** (with a flame icon): SOCH is a set of 3.4 million (as of december 2010) cultural heritage objects harvested from a large number of museums and other local, regional and national cultural heritage... It includes links for HTML, application/rdf+xml, and example/rdf+xml.
- Culture Grid**: About From the website: The Culture Grid is designed to do two things. Firstly, it pulls together info from the thousands of museum, archive and library websites and... It includes links for solr, sru, and oai.
- Flickr - The Commons**: About The key goals of The Commons on Flickr are to firstly show you hidden treasures in the world's public photography archives, and secondly to show how your input and...
- Amsterdam Museum as Linked Open Data in the Europeana Data Model** (with a flame icon): The Amsterdam Museum dataset describes more than 70.000 cultural heritage objects related to the city of Amsterdam described by the museum. The metadata was retrieved from an... It includes links for apisparql, HTML, api/it, and example/rdf+xml.
- British Museum Collection** (with a flame icon): Welcome to this Linked Data and SPARQL service. It provides access to the same collection data available through the Museum's web presented Collection Online, but in a computer-readable format.

mitted to the Linked Data Cloud.

Datasets search

<https://datahub.io/search?q=coronavirus>

Národní katalog otevřených dat (NKOD)

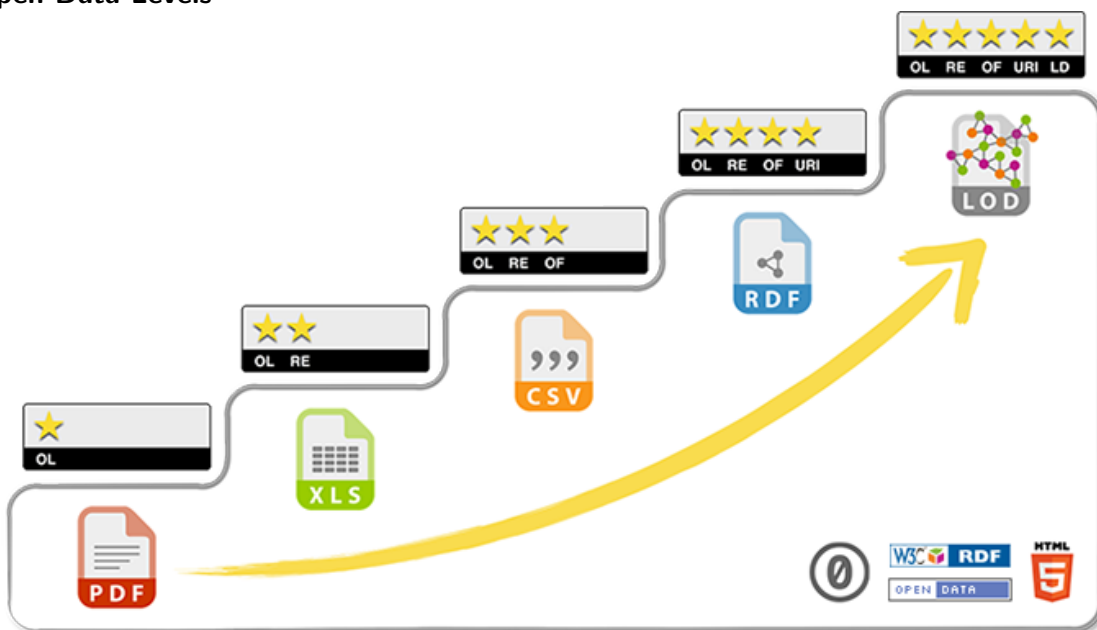
1 Introduction

The screenshot shows a data portal interface. At the top, there are tabs for 'OTEVŘENÁ DATA' and 'DATA'. Below this, there are filters for 'Poskytovatelé (1)' (HLAVNÍ MĚSTO PRAHA (136)), 'Klíčová slova (18)', and 'Formáty (10)'. The main content area displays search results for 'PRAHA', including 'Absolutní výšky budov' and 'Bonita klimatu'. Each result includes a title, a brief description, and a list of available formats (e.g., TIF, Plain text, GeoJSON, Zipped GML, Esri Shape, ZIP).

gov.cz/

https://data.

Open Data Levels



Taken from <http://5stardata.info/cs/>.

Open Data Levels – description

★ Available on the web (whatever format) but with an open licence, to be Open

Data

- ★★ Available as machine-readable structured data (e.g. excel instead of image scan of a table)
- ★★★ All the above, plus – Non-proprietary format (e.g. CSV instead of excel)
- ★★★★ All the above, plus – Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
- ★★★★★ All the above, plus – Link your data to other people’s data to provide context

(Tim Berners-Lee, 2009 – <http://www.w3.org/DesignIssues/LinkedData.html>)

From Open Data to Linked Data

★★★

★★★★

Aircraft (CAA)

s/n	type	operator_ic
1	Boeing 737	1234567
2	Airbus 319	9876543

→ ?

Companies (Business Registry)

company_ic	company_name
1234567	Best Airlines
9876543	Funny Flight School

From Open Data to Linked Data

★★★

★★★★

Aircraft (CAA)

s/n	type	operator_ic
1	Boeing 737	1234567
2	Airbus 319	9876543

→

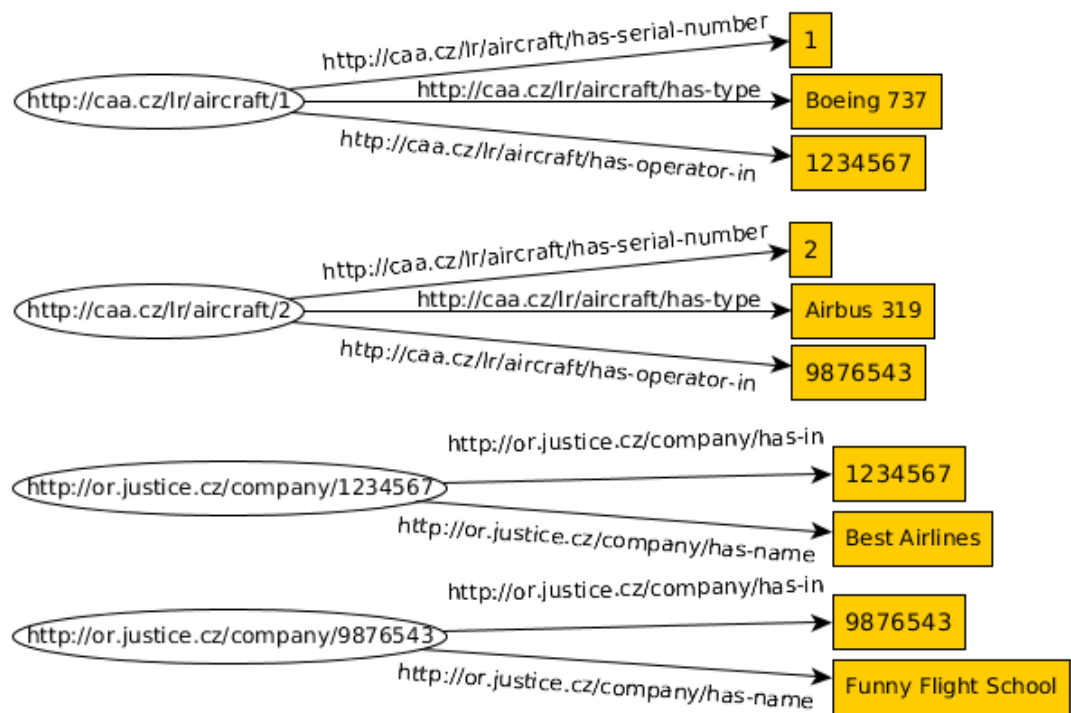


Companies (Business Registry)

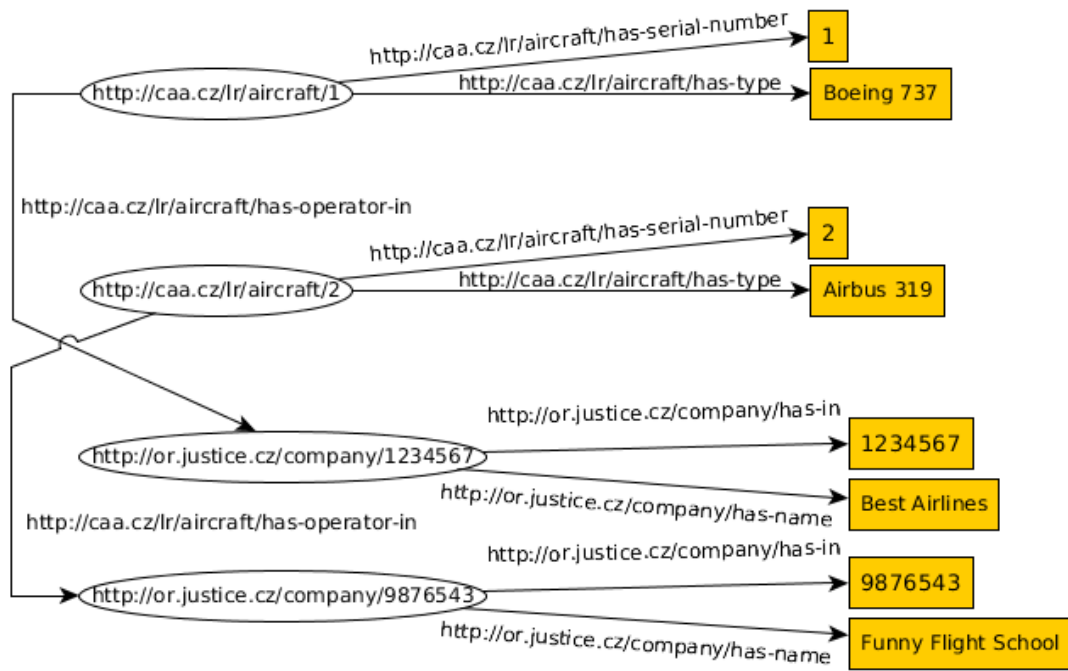
company_ic	company_name
1234567	Best Airlines
9876543	Funny Flight School

From Open Data to Linked Data (4*)

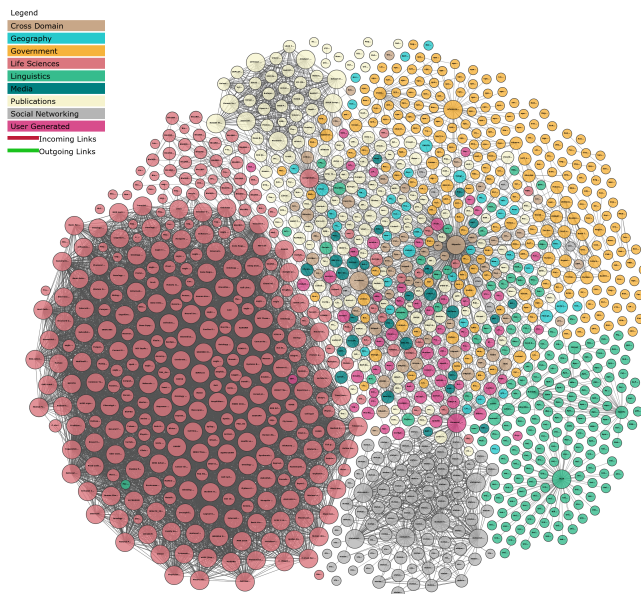
1 Introduction



From Open Data to Linked Data (5*)



Linked Open Data Cloud



<http://lod-cloud.net/>, 2018

1 Introduction

Linked Data vs. Open Data

linked, not open – enterprise data, master data

linked, open – 5* data

not linked, open – typical case in OpenData

not linked, not open – we do not care

1.2.3 Semantic Web Adopters

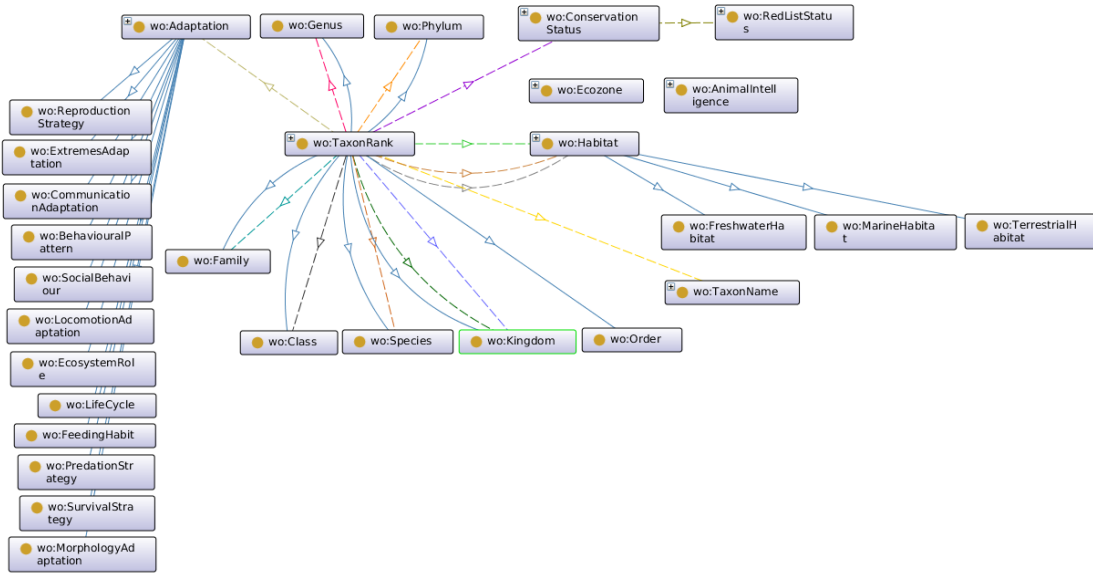
Public Sector

- national governments (e.g in Czechia - DIA - <https://data.gov.cz>)
- EU administration - <https://data.europa.eu/en/publications/datastories/linking-data-dataeuropaeu>
- US administration - <https://data.gov/>
- ...

Private Sector

- Google – *Knowledge Graph* (although they do not name it Semantic web – http://semanticweb.com/google-just-hi-jacked-the-semantic-web-vocabulary_b29092)
- Microsoft – Satori, <http://research.microsoft.com/en-us/projects/trinity/query.aspx>
- Facebook – Open Graph Protocol <http://ogp.me/>
- BBC – various datasets in RDF – <http://www.bbc.co.uk/developer/technology/apis.html>
- Ordnance Survey – geographic datasets in RDF – <http://data.ordnancesurvey.co.uk>

BBC Wildlife Ontology



Ordnance Survey Linked Data Kents Hill, Monkston and Brinklow



Kents Hill, Monkston and Brinklow is a Parish in Milton Keynes.

Objects related to "Kents Hill, Monkston and Brinklow"	
Extent	41649-49
In European Region	South East
Within	Milton Keynes
In District	Milton Keynes
Touches	Walton Broughton Old Woughton Milton Keynes Wavendon
Same As	E04001285

Core facts about "Kents Hill, Monkston and Brinklow"	
Type	Parish
Label	Kents Hill, Monkston and Brinklow
Pref Label	Kents Hill, Monkston and Brinklow
Alt Label	Kents Hill, Monkston and Brinklow CP
Northing	238013.803835
Easting	489602.596729
Lat	52.0333028515
Long	-0.695254366017
Area Code	CPC
Gss Code	E04001285

1 Introduction

Selected Materials

- RDF Primer – <https://www.w3.org/TR/rdf11-primer/>
- SPARQL Query Language Spec – <https://www.w3.org/TR/2013/REC-sparql11-quer>
- OWL Primer – <https://www.w3.org/TR/owl2-primer/>
- SKOS Primer – <https://www.w3.org/TR/skos-primer/>
- Description Logic Reasoning – P. Křemen, *Ontologie a Deskripční logiky*. In *Umělá inteligence VI.*, Academia, 2013.
- Linked Data – <http://linkeddata.org>
- Tutorial on RDF/OWL – <https://www.obitko.com/tutorials/ontologies-semantic>