Introduction

Petr Křemen

petr.kremen@fel.cvut.cz

September 24, 2020

Outline

1. Why this Course?
2. Overview of Ontologies
3. Use-case: Data Integration
4. Semantic Web
   - Semantic Web Adopters
5. Linked Data
6. Use-case: Open Data

1 Why this Course?

2 Overview of Ontologies

3 Use-case: Data Integration

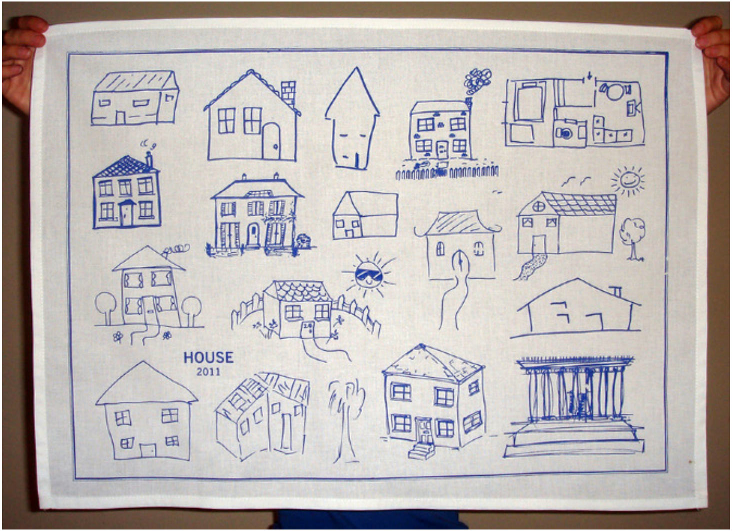
4 Semantic Web
- Semantic Web Adopters

5 Linked Data

6 Use-case: Open Data

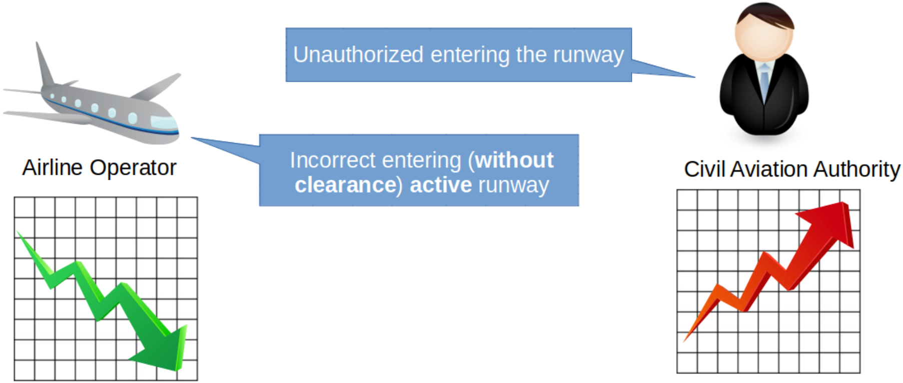
Why this Course?

What is a house ?



Why to care ?

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



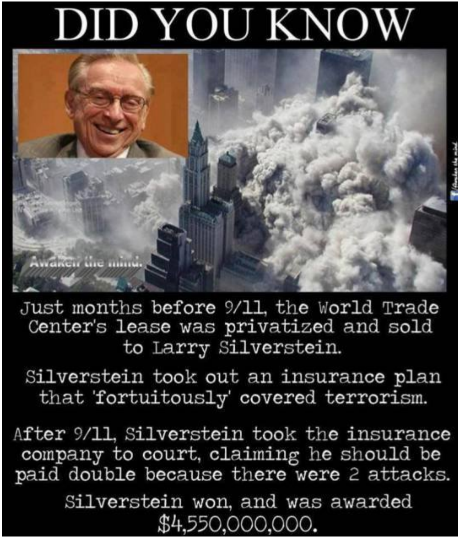
Unauthorized entering the runway

Airline Operator

Incorrect entering (**without clearance**) **active** runway

Civil Aviation Authority

Why to care ?



DID YOU KNOW

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.

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

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

Knowledge Management

9/11 ... matter of billions of USD

About ontologies

Ontologies

are **formal specifications of conceptualization**.

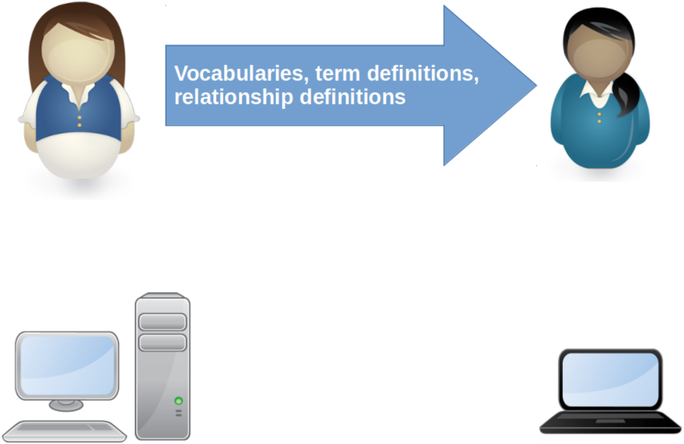
Ontologies help to stabilize the knowledge, to share meaning both among computers and among people. Use-cases include

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

1. Why this Course?
2. Overview of Ontologies
3. Use-case: Data Integration
4. Semantic Web
   - Semantic Web Adopters
5. Linked Data
6. Use-case: Open Data

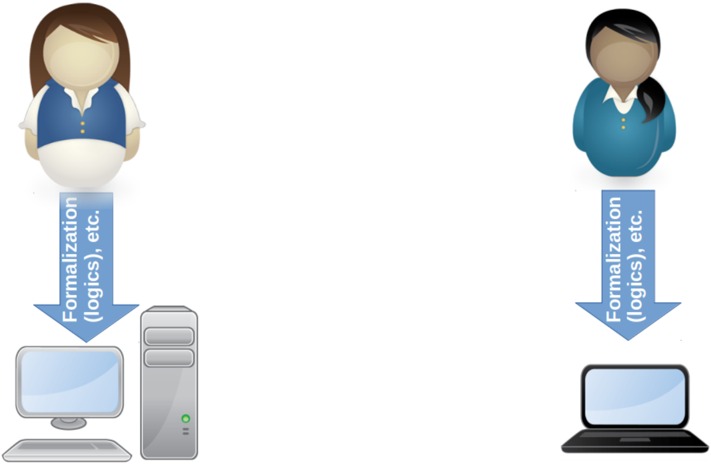
Overview of Ontologies

First, People Need to Understand Each Other

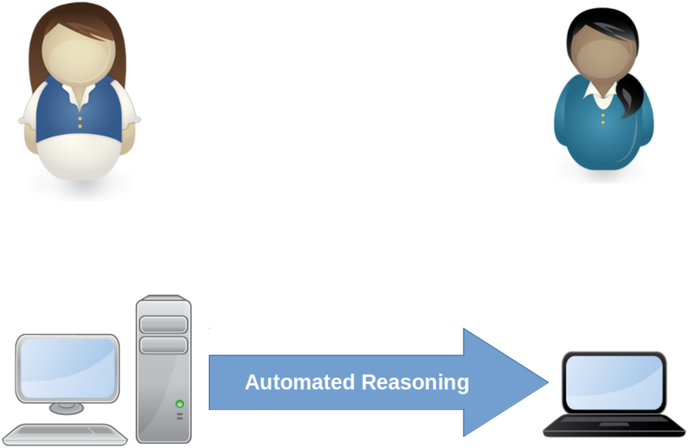


Vocabularies, term definitions, relationship definitions

Second, People Need to Explain Things to Computers



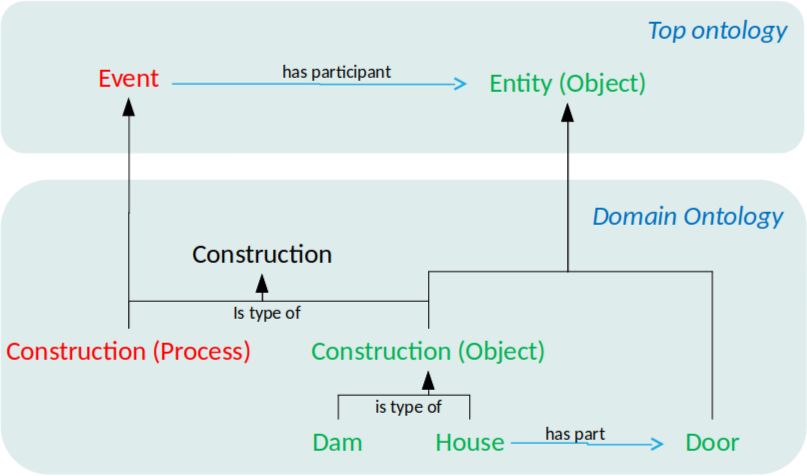
Third, Computers Can Understand One Another



Automated Reasoning

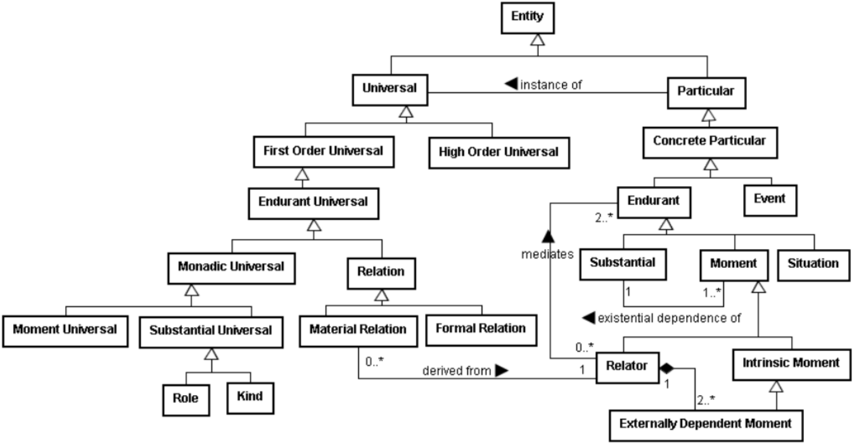
Solution = Ontology

Explicit Conceptualization of Shared Meaning



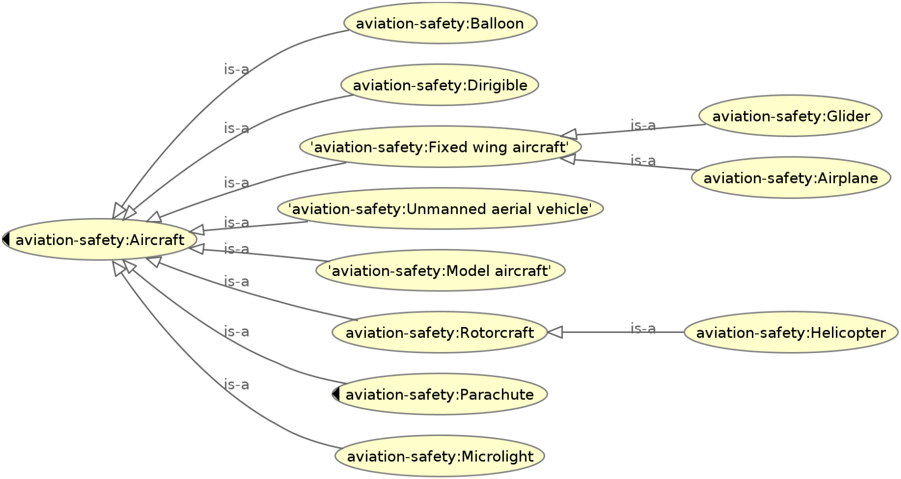
Example Top-Level Ontology

Small part of Unified Foundational Ontology (UFO)



Example Ontology Hierarchy

Each helicopter is also an aircraft.


aviation-safety:Balloon
aviation-safety:Dirigible
'aviation-safety:Fixed wing aircraft'
aviation-safety:Glider
aviation-safety:Airplane
'aviation-safety:Unmanned aerial vehicle'
aviation-safety:Aircraft
'aviation-safety:Model aircraft'
aviation-safety:Rotorcraft
aviation-safety:Helicopter
aviation-safety:Parachute
aviation-safety:Microlight
is-a
is-a
is-a
is-a
is-a
is-a
is-a
is-a
is-a
is-a
is-a

Ontologies ≠ Taxonomies

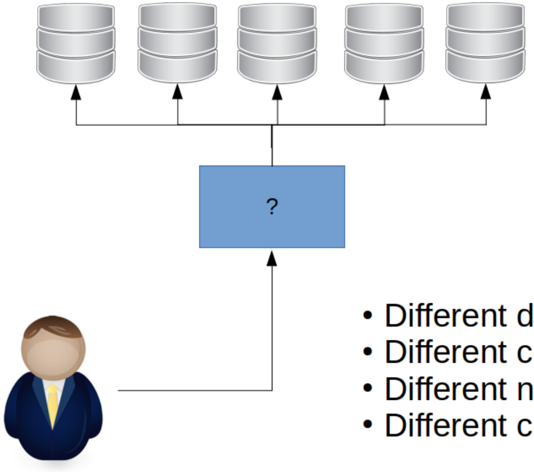
Taxonomies = just a single type of relationship.

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

1. Why this Course?
2. Overview of Ontologies
3. Use-case: Data Integration
4. Semantic Web
   - Semantic Web Adopters
5. Linked Data
6. Use-case: Open Data

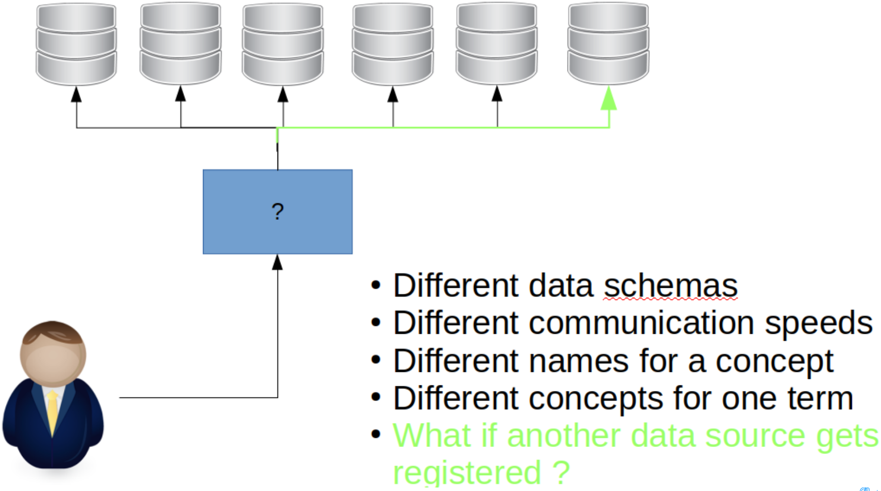
Use-case: Data Integration

Data Integration Scenario



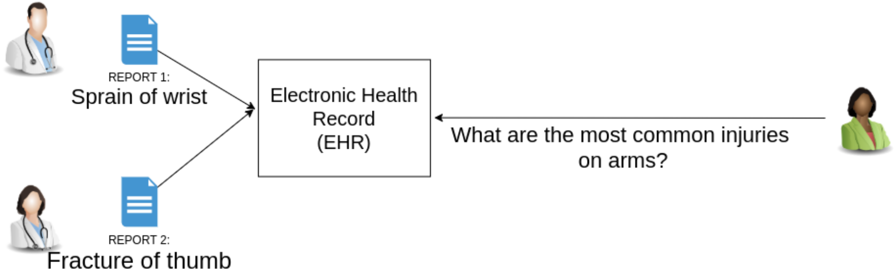
- Different data schemas
- Different communication speeds
- Different names for a concept
- Different concepts for one term

Data Integration Scenario



- Different data schemas
- Different communication speeds
- Different names for a concept
- Different concepts for one term
- What if another data source gets registered ?

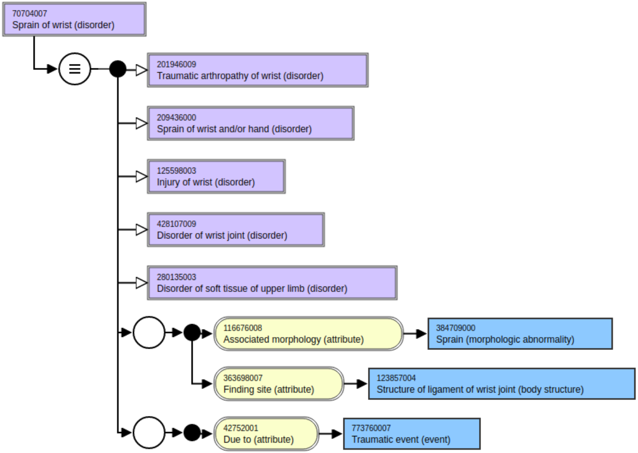
Use-case – HealthCare Data Integration


REPORT 1:
Sprain of wrist
REPORT 2:
Fracture of thumb
Electronic Health Record (EHR)
What are the most common injuries on arms?

SNOMED-CT

Systematized Nomenclature of Medicine - Clinical Terms

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



70704007
Sprain of wrist (disorder)

≡

201946009
Traumatic arthropathy of wrist (disorder)

209436000
Sprain of wrist and/or hand (disorder)

125598003
Injury of wrist (disorder)

428107009
Disorder of wrist joint (disorder)

280135003
Disorder of soft tissue of upper limb (disorder)

116676008
Associated morphology (attribute)

384709000
Sprain (morphologic abnormality)

363698007
Finding site (attribute)

123857004
Structure of ligament of wrist joint (body structure)

42752001
Due to (attribute)

773760007
Traumatic event (event)

SNOMED-CT

Systematized Nomenclature of Medicine - Clinical Terms

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

1. Why this Course?
2. Overview of Ontologies
3. Use-case: Data Integration
4. Semantic Web
   - Semantic Web Adopters
5. Linked Data
6. Use-case: Open Data

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.

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 ?

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),

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,

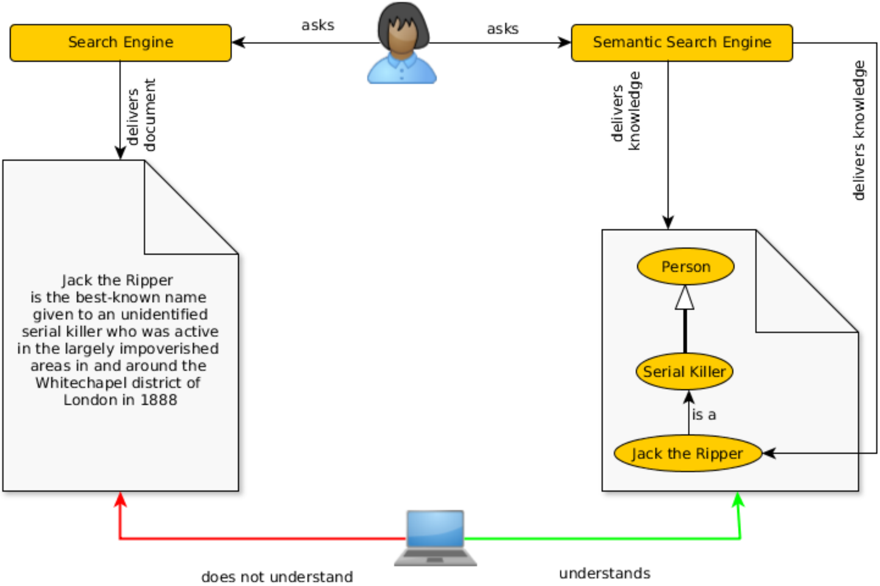
Current Web vs. Semantic Web

- 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

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



Jack the Ripper
is the best-known name
given to an unidentified
serial killer who was active
in the largely impoverished
areas in and around the
Whitechapel district of
London in 1888

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`

Idea of Semantic Web

- W3C web page - `http://www.w3.org/2001/sw`
- The data format will be either RDF(S) or OWL,

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,

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 adressing 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

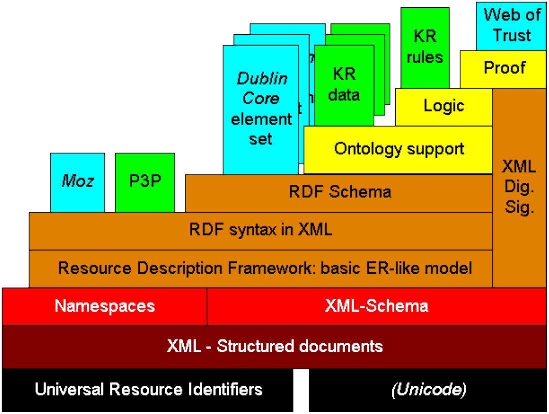
The semantic web inference must take into account that we handle *incomplete knowledge*.

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.

Semantic Web Adopters

1 Why this Course?

2 Overview of Ontologies

3 Use-case: Data Integration

4 Semantic Web
- Semantic Web Adopters

5 Linked Data

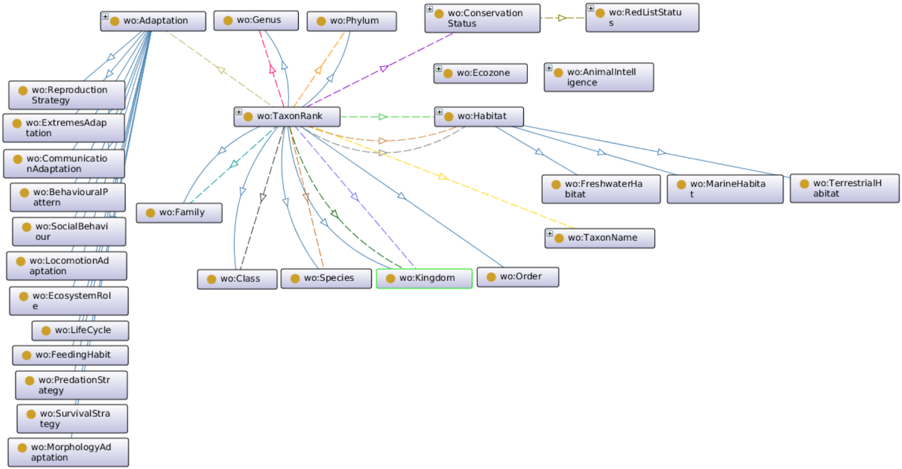
6 Use-case: Open Data

Who is Using Semantic Web Technologies

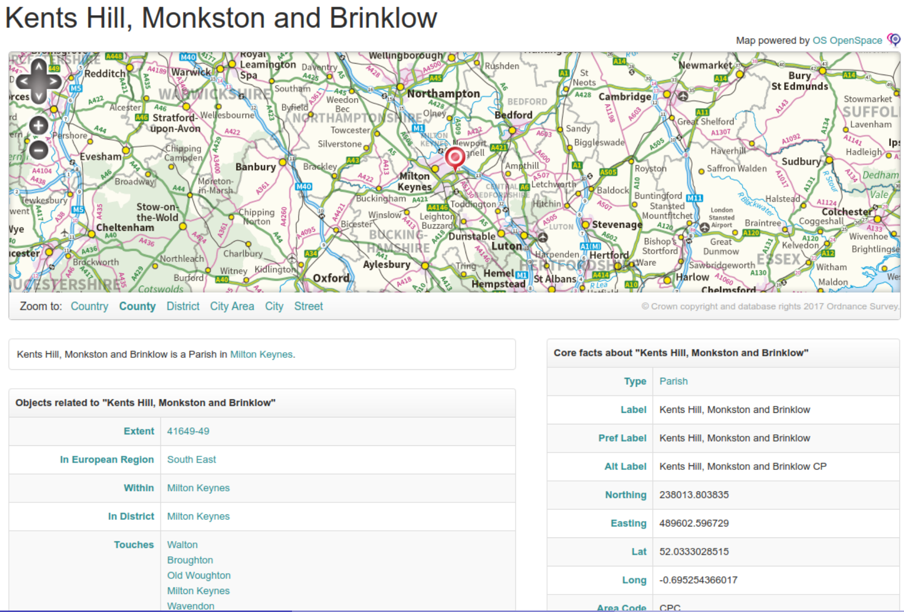
Let's name a few:

- 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

Map powered by OS OpenSpace

Zoom to: Country County District City Area City Street

© Crown copyright and database rights 2017 Ordnance Survey.

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

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

1 Why this Course?

2 Overview of Ontologies

3 Use-case: Data Integration

4 Semantic Web
- Semantic Web Adopters

5 Linked Data

6 Use-case: Open Data

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

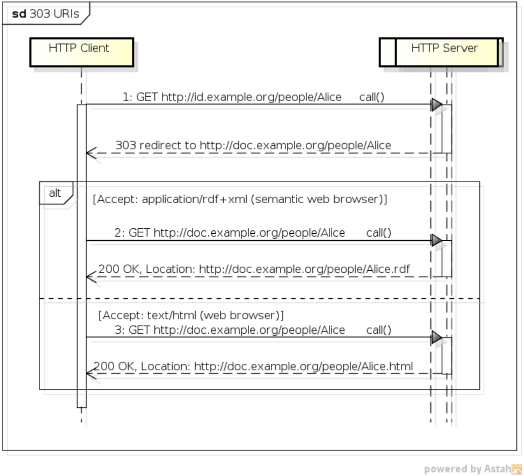
```
@prefix people: <http://example.com/people/>
people:John people:likes people:Mary
```

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.

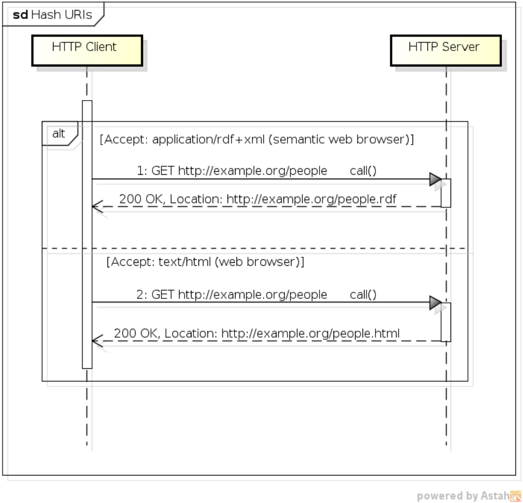
303 URIs

- 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.


sd 303 URIs
HTTP Client
HTTP Server
1: GET http://id.example.org/people/Alice call()
303 redirect to http://doc.example.org/people/Alice
alt
[Accept: application/rdf+xml (semantic web browser)]
2: GET http://doc.example.org/people/Alice call()
200 OK, Location: http://doc.example.org/people/Alice.rdf
[Accept: text/html (web browser)]
3: GET http://doc.example.org/people/Alice call()
200 OK, Location: http://doc.example.org/people/Alice.html
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.


sd Hash URIs
HTTP Client
HTTP Server
alt
[Accept: application/rdf+xml (semantic web browser)]
1: GET http://example.org/people call()
200 OK, Location: http://example.org/people.rdf
[Accept: text/html (web browser)]
2: GET http://example.org/people call()
200 OK, Location: http://example.org/people.html
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 performace.

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 Platforms

Pubby is a simple Linked Data publication server connectable to SPARQL endpoints,

Callimachus is an application server for linked data applications. To be explored in the tutorials,

Marmotta is a platform for publishing Linked Data (contributed from Linked Media Framework),

D2R is a platform for publishing relational database data in the form of Linked Data.

1 Why this Course?

2 Overview of Ontologies

3 Use-case: Data Integration

4 Semantic Web
- Semantic Web Adopters

5 Linked Data

6 Use-case: Open Data

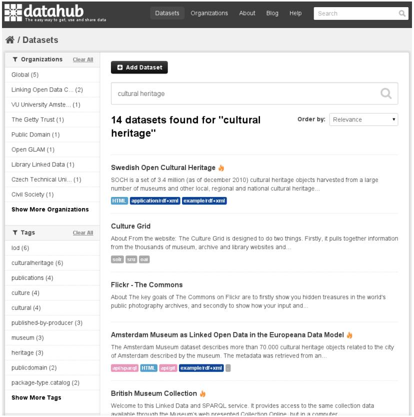
Use-case: Open Data

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 submitted to the Linked Data Cloud.



14 datasets found for "cultural heritage"

Swedish Open Cultural Heritage

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...

HTML application/rdf+xml example/rdf+xml

Culture Grid

About From the website: The Culture Grid is designed to do two things. Firstly, it pulls together information from the thousands of museum, archive and library websites and...

solr sru 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

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...

api/sparql HTML api/git example/rdf+xml -

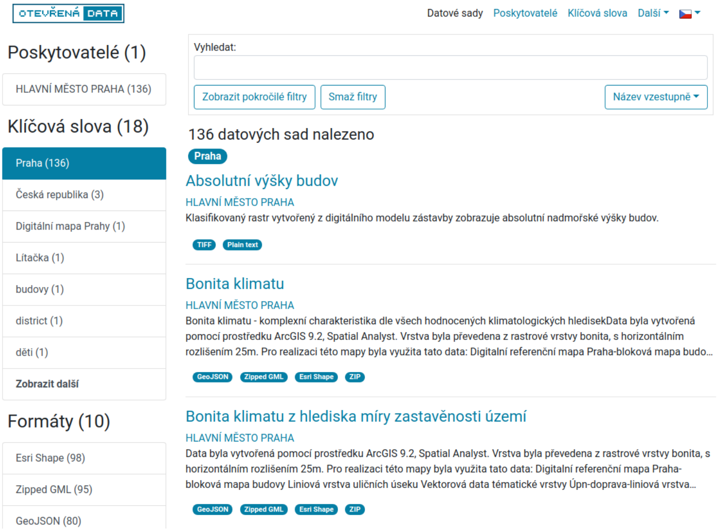
British Museum Collection

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

Datasets search

`https://datahub.io/search?q=coronavirus`

Národní katalog otevřených dat (NKOD)



Poskytovatelé (1)

HLAVNÍ MĚSTO PRAHA (136)

Klíčová slova (18)

- Praha (136)
- Česká republika (3)
- Digitální mapa Prahy (1)
- Lítačka (1)
- budovy (1)
- district (1)
- děti (1)
- **Zobrazit další**

Formáty (10)

- Esri Shape (98)
- Zipped GML (95)
- GeoJSON (80)

Vyhledat:

136 datových sad nalezeno

Praha

Absolutní výšky budov

HLAVNÍ MĚSTO PRAHA

Klasifikovaný rastr vytvořený z digitálního modelu zástavby zobrazuje absolutní nadmořské výšky budov.

TIFF Plain text

Bonita klimatu

HLAVNÍ MĚSTO PRAHA

Bonita klimatu - komplexní charakteristika dle všech hodnocených klimatologických hlediskData byla vytvořená pomocí prostředku ArcGIS 9.2, Spatial Analyst. Vrstva byla převedena z rastrové vrstvy bonita, s horizontálním rozlišením 25m. Pro realizaci této mapy byla využita tato data: Digitalní referenční mapa Praha-bloková mapa budo...

GeoJSON Zipped GML Esri Shape ZIP

Bonita klimatu z hlediska míry zastavěnosti území

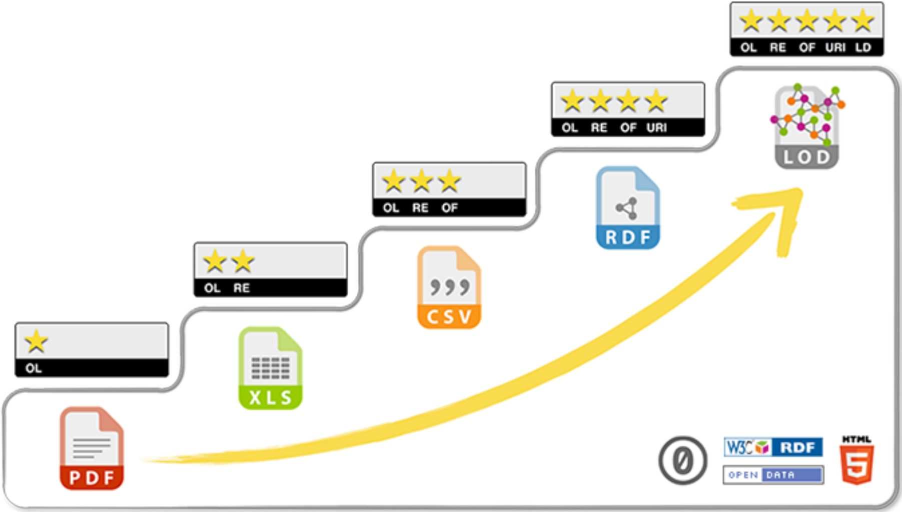
HLAVNÍ MĚSTO PRAHA

Data byla vytvořená pomocí prostředku ArcGIS 9.2, Spatial Analyst. Vrstva byla převedena z rastrové vrstvy bonita, s horizontálním rozlišením 25m. Pro realizaci této mapy byla využita tato data: Digitalní referenční mapa Praha-bloková mapa budovy Liniová vrstva uličních úseku Vektorová data tématické vrstvy Úpn-doprava-liniová vrstva...

GeoJSON Zipped GML Esri Shape ZIP

`https://data.gov.cz/`

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

⋆⋆⋆ ⋆⋆⋆⋆⋆

Aircrafts (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

⋆⋆⋆ ⋆⋆⋆⋆

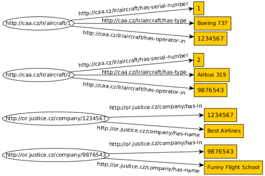
Aircrafts (CAA)

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

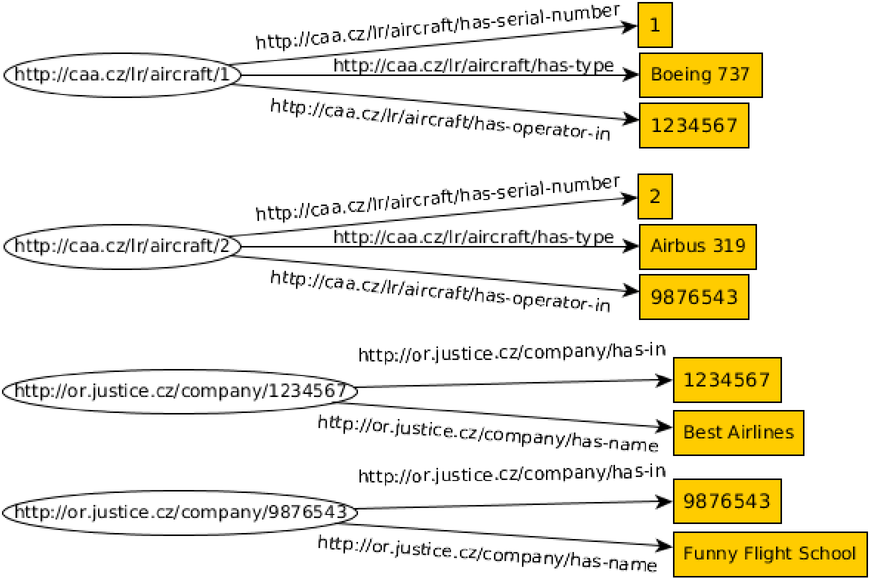
Companies (Business Registry)

company_ic	company_name
1234567	Best Airlines
9876543	Funny Flight School

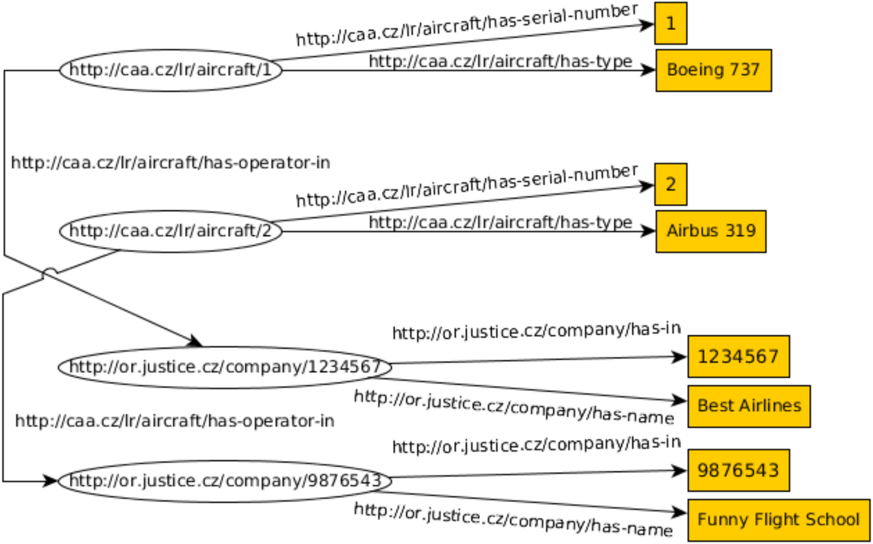
→


http://caa.cz/lr/aircraft/1
http://caa.cz/lr/aircraft/has-serial-number
1
http://caa.cz/lr/aircraft/has-type
Boeing 737
http://caa.cz/lr/aircraft/has-operator-in
1234567
http://caa.cz/lr/aircraft/2
http://caa.cz/lr/aircraft/has-serial-number
2
http://caa.cz/lr/aircraft/has-type
Airbus 319
http://caa.cz/lr/aircraft/has-operator-in
9876543
http://or.justice.cz/company/1234567
http://or.justice.cz/company/has-in
1234567
http://or.justice.cz/company/has-name
Best Airlines
http://or.justice.cz/company/9876543
http://or.justice.cz/company/has-in
9876543
http://or.justice.cz/company/has-name
Funny Flight School

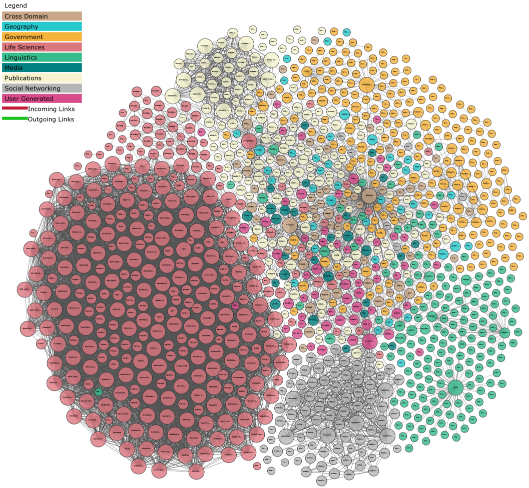
From Open Data to Linked Data (4*)


http://caa.cz/lr/aircraft/1
http://caa.cz/lr/aircraft/has-serial-number
1
http://caa.cz/lr/aircraft/has-type
Boeing 737
http://caa.cz/lr/aircraft/has-operator-in
1234567
http://caa.cz/lr/aircraft/2
http://caa.cz/lr/aircraft/has-serial-number
2
http://caa.cz/lr/aircraft/has-type
Airbus 319
http://caa.cz/lr/aircraft/has-operator-in
9876543
http://or.justice.cz/company/1234567
http://or.justice.cz/company/has-in
1234567
http://or.justice.cz/company/has-name
Best Airlines
http://or.justice.cz/company/9876543
http://or.justice.cz/company/has-in
9876543
http://or.justice.cz/company/has-name
Funny Flight School

From Open Data to Linked Data (5*)


http://caa.cz/lr/aircraft/has-serial-number
1
http://caa.cz/lr/aircraft/1
http://caa.cz/lr/aircraft/has-type
Boeing 737
http://caa.cz/lr/aircraft/has-operator-in
http://caa.cz/lr/aircraft/has-serial-number
2
http://caa.cz/lr/aircraft/2
http://caa.cz/lr/aircraft/has-type
Airbus 319
http://or.justice.cz/company/has-in
http://or.justice.cz/company/1234567
1234567
http://or.justice.cz/company/has-name
Best Airlines
http://caa.cz/lr/aircraft/has-operator-in
http://or.justice.cz/company/has-in
http://or.justice.cz/company/9876543
9876543
http://or.justice.cz/company/has-name
Funny Flight School

Linked Open Data Cloud



`http://lod-cloud.net/`,2018

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

Selected Materials

- OSW pages – `https://cw.fel.cvut.cz/wiki/courses/osw`
- RDF Primer – `https://www.w3.org/TR/rdf11-primer/`
- SPARQL Query Language Spec – `https://www.w3.org/TR/2013/REC-sparql11-query-20130321/`
- 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`
- Nice supplementary tutorial on RDF/OWL – `https://www.obitko.com/tutorials/ontologies-semantic-web/`