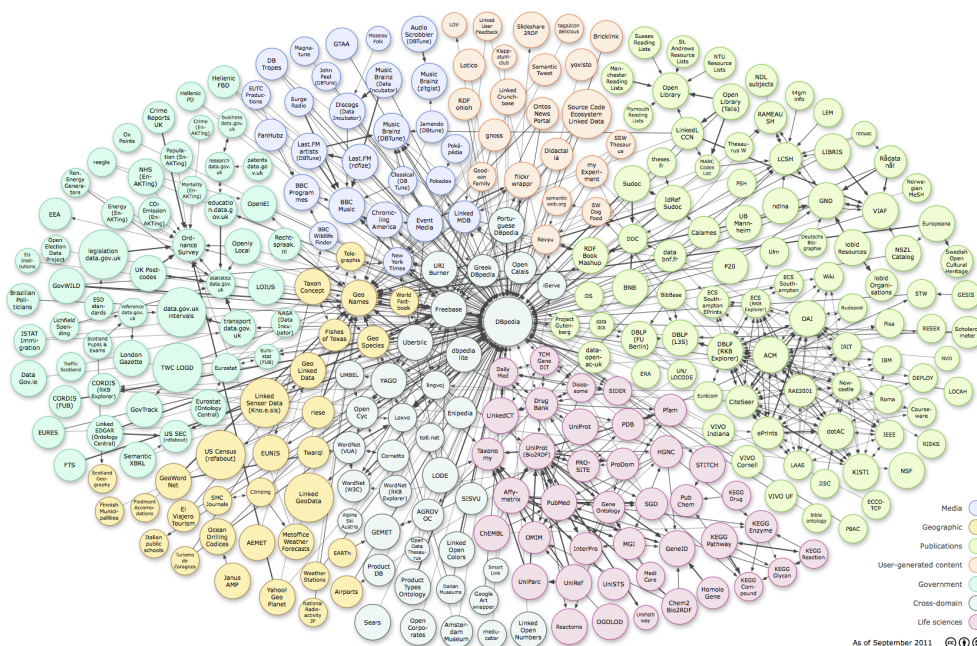


# 1 Managing Semantic Data

## 1.1 Linked Data

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

### Linking Open Data cloud diagram



... by R.Cygniak and A.Jentzsch. <http://lod-cloud.net/>, 2011

### Statistics for the Linked Data Cloud

Domain	Number of datasets	Triples
Media	25	1,841,852,061
Geographic	31	6,145,532,484
Government	49	13,315,009,400
Publications	87	2,950,720,693
Cross-domain	41	4,184,635,715
Life sciences	41	3,036,336,004
User-generated content	20	134,127,413
	<b>295</b>	<b>31,634,213,770</b>

(in 2011)

Online formalized statistics are available at <http://stats.lod2.eu>.

### 1.1.1 Core Linked Data

#### Classical Web vs. Semantic Web

- semantic web (RDF) links things, not just documents,

#### Example

RDF connects a person described within one document with its friends described in another document, while HTML links only these two documents.

- semantic web (hyper)links are typed

#### Example

RDF tells what kind of relationship between the two persons is (e.g. is-friend-of), while HTML hyper-links do not.

#### 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
```

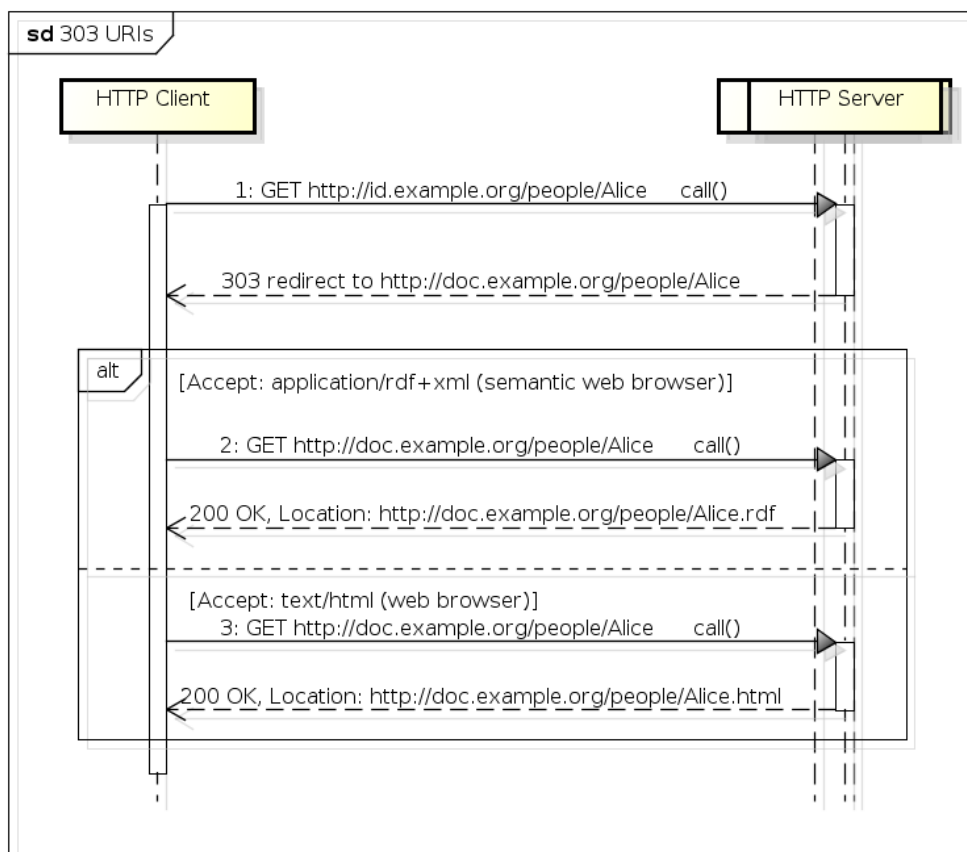
Is `http://example.com/people/Mary` a web document or a resource ? (Consider semantic consequences of each option).

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.

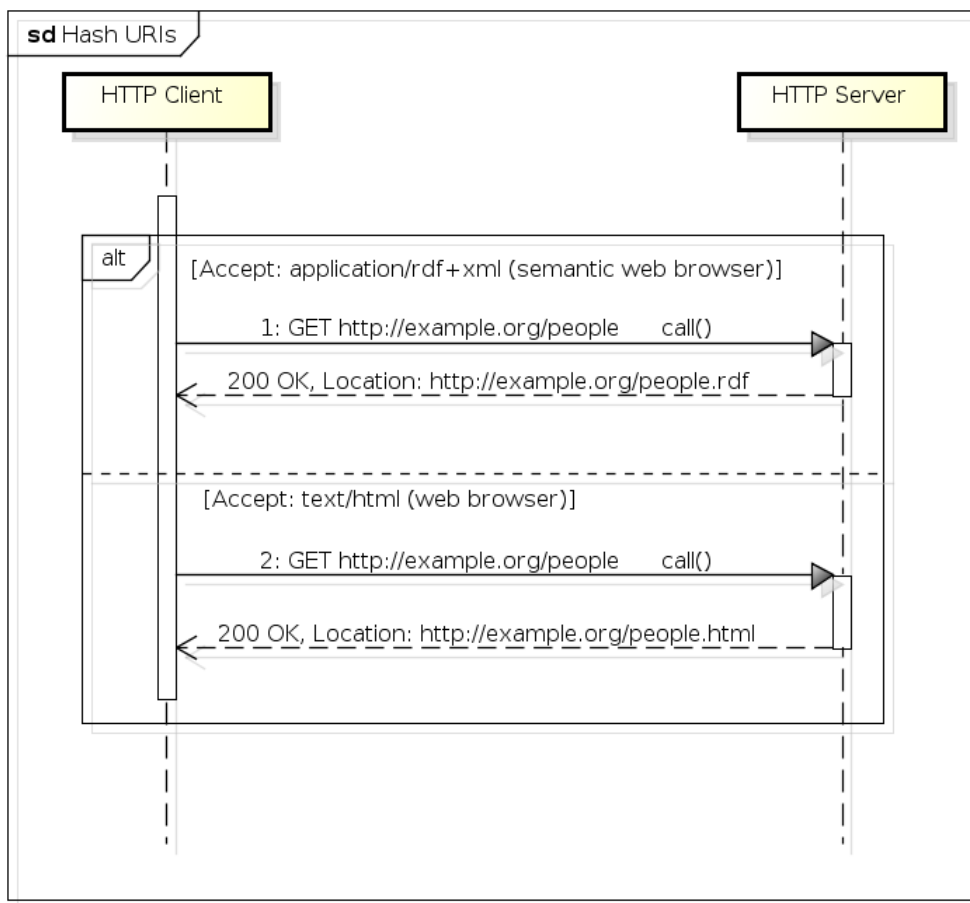
## 1 Managing Semantic Data



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.

### 1.1.2 Creating Linked Data

#### Dereferencability of URIs

To make a URI dereferencable there are a few options:

**Custom HTTP server** – run a generic HTTP server (like Apache2) hosting your domain and configure the corresponding responses

**Linked Data Platform** – use existing Linked Data Platform

- for publishing RDF data backed by a SPARQL endpoint (Pubby)
- for publishing RDBMS data as Linked Data (D2R)
- for creating Linked Data Applications (Linked Data Platforms [Arwe:14:LDP], like Callimachus)

When designing a URI scheme, you should consider making them as Cool URIs [Sauermann:08:CUS] PURL (<http://purl.org>) might be used to ensure stability of the identifier.

### Developing Custom Linked Data

1. Find relevant vocabularies/ontologies using existing services – e.g. DBPedia, Watson, Sindice (see section 1.1.3)
2. Extend these vocabularies/ontologies by introducing new resources in the namespace under your control.
3. Develop the dataset in RDF model (create the data/publish existing RDBMS data, etc.)
4. Intelink the dataset to existing data (search suitable datasets using similar techniques as in step 1)
5. Publish the Data using one of the options discussed in section 1.1.4.

### Linking other data

There are a few types of links to other datasets.

**Relationship links** connect two individuals/objects (like people, places, animals) with some relationship (like knows, bornIn) from different datasets.

**Identity links** are placed to say that two individuals/objects are identical (e.g. describing one person in two datasets under different identities)

**Vocabulary links** point from data to the definitions (schema/ontology)

### Suitable Vocabularies

**FOAF** (Friend Of A Friend) <http://xmlns.com/foaf/0.1/> – linking people and information about them

**VCard** (<http://www.w3.org/2006/vcard/ns#>) – describing people and organizations from the business perspective

**Relationship** (<http://purl.org/vocab/relationship>) – describing relationships between people

**Basic GEO** (<http://www.w3.org/2003/01/geo>) – describing basic spatial locations (more in lecture on Semantic GIS)

**Dublin Core** (<http://dublincore.org/documents/dcmi-terms>) – events and annotation of documents

**Event Ontology** (<http://purl.org/NET/c4dm/event.owl>) – temporal events

**GoodRelations** (<http://purl.org/goodrelations/v1#>) – E-commerce schema for describing products and their offers

**VOID** (<http://rdfs.org/ns/void#>) – description of a dataset

**DOAP** (Description Of A Project) (<http://usefulinc.com/ns/doap#>) – description of open-source projects

**Sitemaps** (<http://www.sitemaps.org/schemas/sitemap/0.9>) – semantic sitemaps vocabulary

### 1.1.3 Tools for Getting and Interlinking Data

#### Apache Any23

Apache Any23 (<http://any23.apache.org/>) is an online service<sup>1</sup> for extracting structured data (RDF/XML, CSV) into arbitrary RDF format.

**inputs** – e.g. RDF/XML, Turtle, N3, RDFa, CSV

**output** – any RDF format

#### Example

<http://any23.org/any23/?format=ttl&uri=http://dbpedia.org/resource/Praha> <http://any23.org/any23/best/https://www.linkedin.com/in/jaracimrman>

#### Task

Try to extract your LinkedIn/Twitter profile information into Turtle.

#### LDSpider

- LDSpider (<https://code.google.com/p/ldspider>) is a linked-data crawler (available as CLI and library). It takes an initial RDF resource and returns an RDF graph representing the resource by dereferencing the URIs to the configurable depth.
- More in tutorials.

#### Watson

Watson (<http://watson.kmi.open.ac.uk>) is a RDF document indexing and search engine .

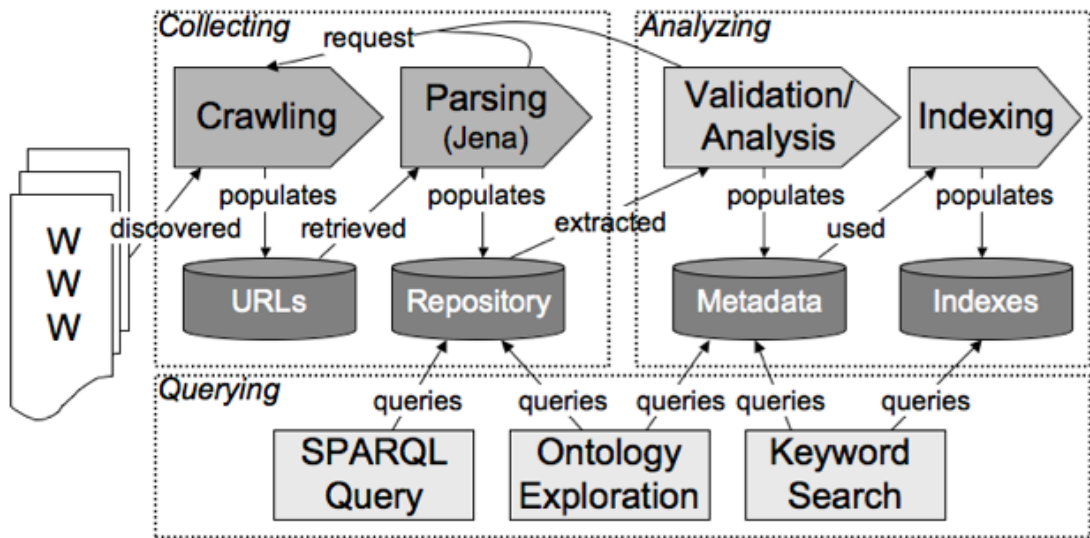
- input – a set of keywords

---

<sup>1</sup>also available as CLI and library

## 1 Managing Semantic Data

- output – a set of RDF documents containing the keywords in their textual fields (e.g. `rdf:label`)



### Sindice

Sindice (<http://sindice.com>) is the Semantic Web Index indexes semantic web.

### Sig.ma

Sig.ma (<http://sig.ma>) is a Linked Data mashup tool

- input – a list of keywords
- output – a pretty-printed RDF graph subject of which is a resource “labeled by the input list of keywords”.



### 1.1.4 Publishing Linked Open Data

#### Linked Data Publishing Checklist

- Does your data set links to other data sets?
- Do you provide provenance metadata? (e.g. VOID descriptions)
- Do you provide licensing metadata?
- Do you use terms from widely deployed vocabularies?
- Are the URIs of proprietary vocabulary terms dereferenceable?
- Do you map proprietary vocabulary terms to other vocabularies?
- Do you provide data set-level metadata?
- Do you refer to additional access methods?

#### Five Stars

- ★ 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

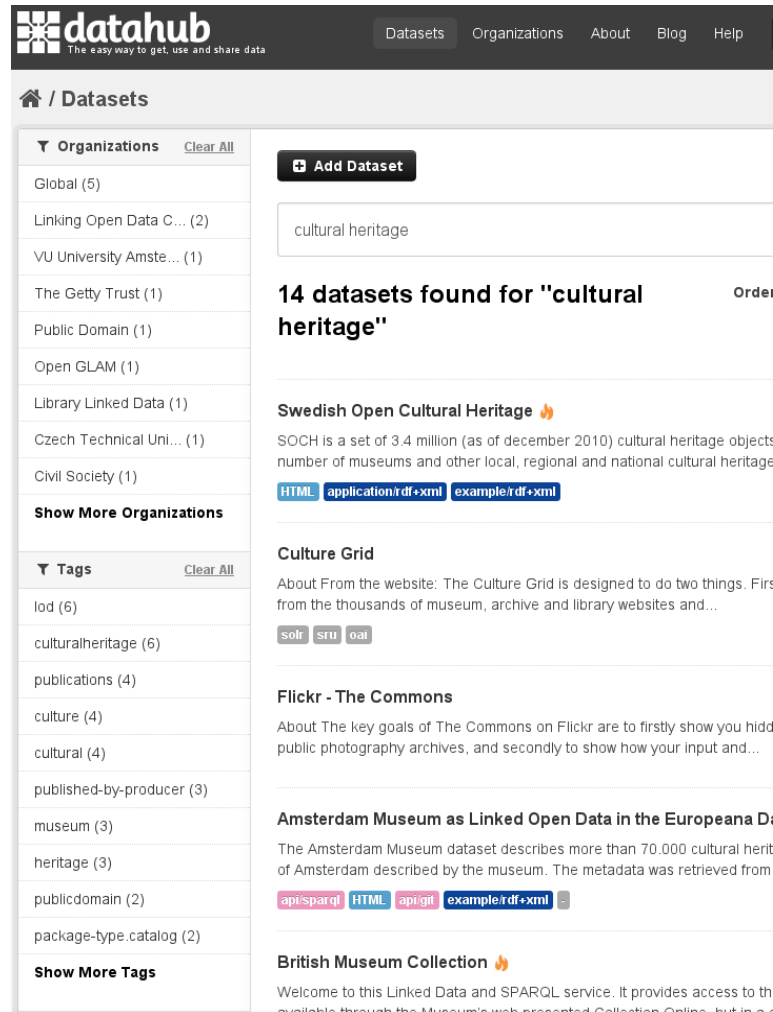
(Tim Berners-Lee, 2009 – <http://www.w3.org/DesignIssues/LinkedData.html>)

#### CKAN and DataHub

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

## 1 Managing Semantic Data

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". Below the navigation bar, the page title is "/ Datasets". On the left side, there are two filter sections: "Organizations" and "Tags". The "Organizations" section lists various organizations with their respective dataset counts, such as "Global (5)", "Linking Open Data C... (2)", and "VU University Amste... (1)". The "Tags" section lists various tags with their respective dataset counts, such as "lod (6)", "culturalheritage (6)", and "publications (4)". On the right side, there is a search bar containing the text "cultural heritage". Below the search bar, there is a button labeled "Add Dataset". The search results section displays "14 datasets found for 'cultural heritage'". The first result is "Swedish Open Cultural Heritage" with a flame icon. The description for this result states: "SOCH is a set of 3.4 million (as of december 2010) cultural heritage objects number of museums and other local, regional and national cultural heritage". Below the description, there are three buttons: "HTML", "application/rdf+xml", and "example/rdf+xml". The second result is "Culture Grid" with a description: "About From the website: The Culture Grid is designed to do two things. First from the thousands of museum, archive and library websites and...". Below the description, there are three buttons: "solr", "stu", and "oai". The third result is "Flickr - The Commons" with a description: "About The key goals of The Commons on Flickr are to firstly show you hidd public photography archives, and secondly to show how your input and...". The fourth result is "Amsterdam Museum as Linked Open Data in the Europeana D" with a description: "The Amsterdam Museum dataset describes more than 70.000 cultural herit of Amsterdam described by the museum. The metadata was retrieved from". Below the description, there are four buttons: "api/sparql", "HTML", "api/glt", and "example/rdf+xml". The fifth result is "British Museum Collection" with a flame icon and a description: "Welcome to this Linked Data and SPARQL service. It provides access to th available through the Museum's web presented Collection Online. but in a".

mitted to the Linked Data Cloud.

### Datasets search

<http://datahub.io/dataset?q=cultural+heritage>

### 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.

## 1.2 Open Data

### 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

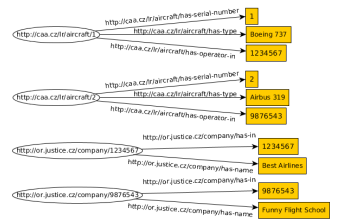
\*\*\*

\*\*\*

Aircrafts (CAA)

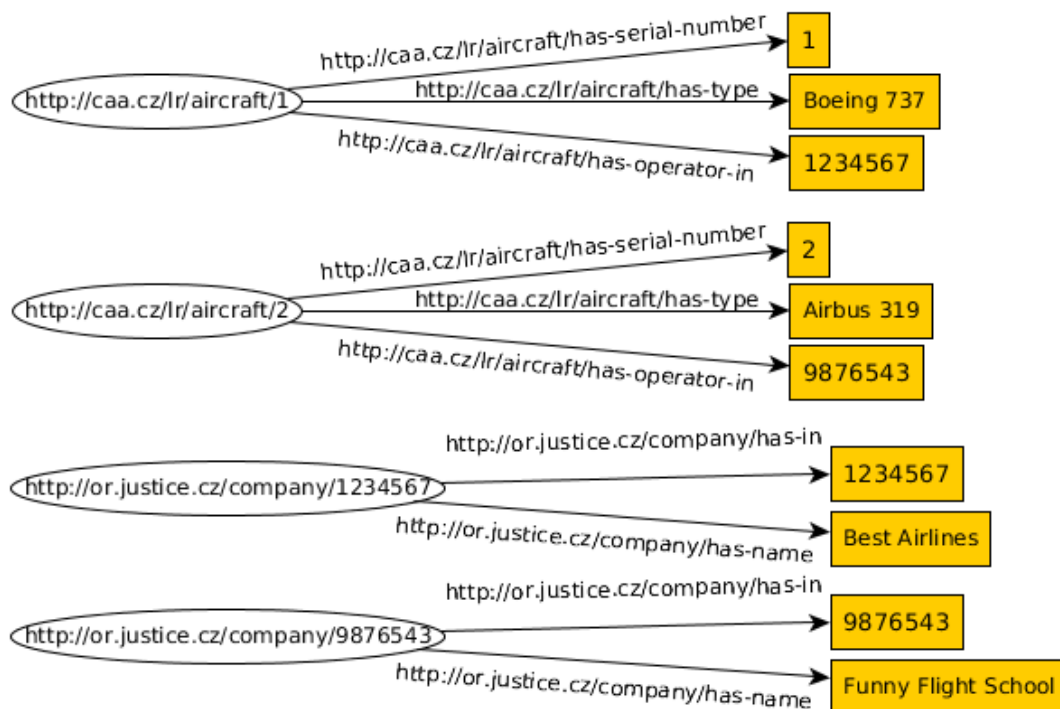
s/n	type	operator_ic
1	Boeing 737	1234567
2	Airbus 319	9876543

→

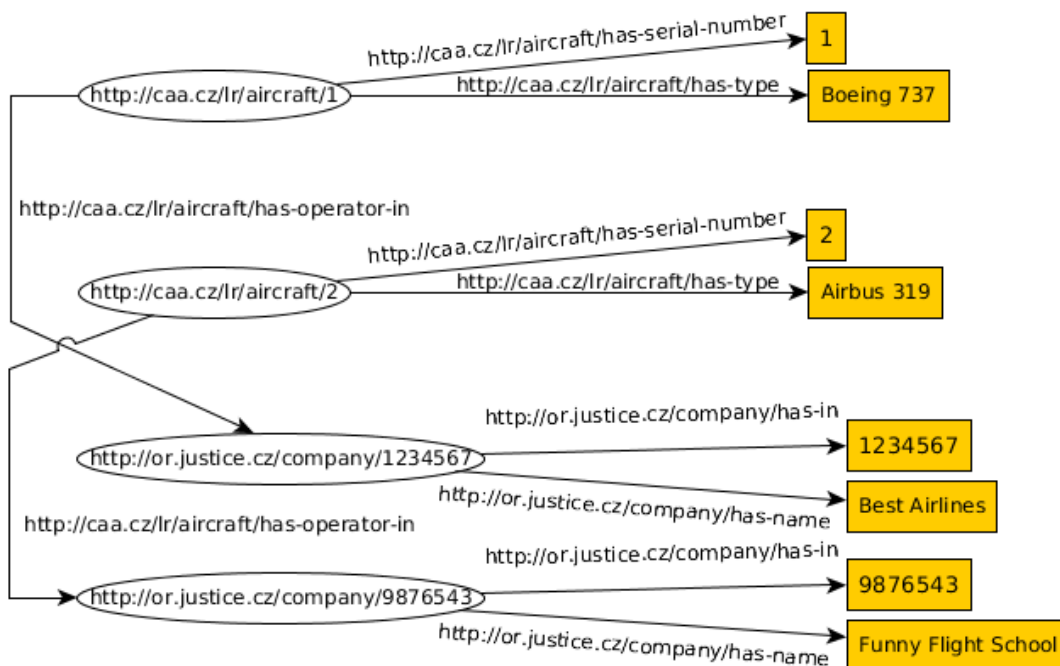


### From Open Data to Linked Data (4\*)

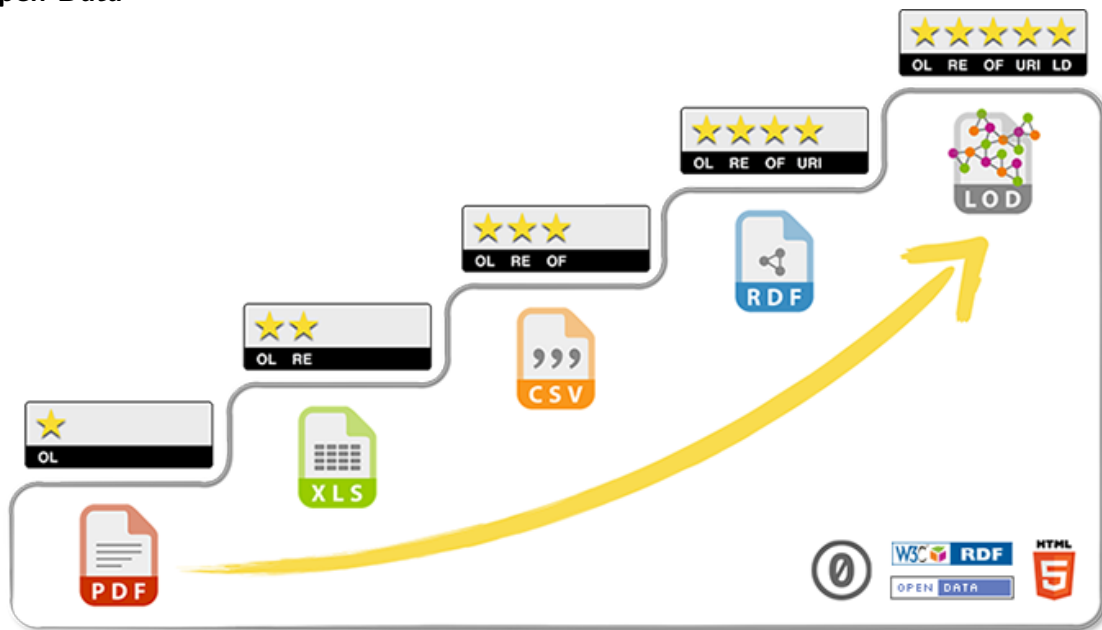
# 1 Managing Semantic Data



## From Open Data to Linked Data (5\*)



## Open Data



Taken

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

### 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

### 1.2.1 Licensing Open Data

#### Open Definition (OD)

Choosing an appropriate license is a crucial point influencing possibilities of future reuse of your data as well as defining your responsibility for the data. Linked data can be used for enterprise (closed) data, as well as open data. Let's discuss licensing of the latter.

**Open Definition** – A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike. – cit. from <http://opendefinition.org>

#### Selected OD-Conformant Creative Commons Licenses

The following licenses apply to both *data* (in the sense of a full database), as well as their *content* (in the sense of particular single statements from these databases).

1 Managing Semantic Data

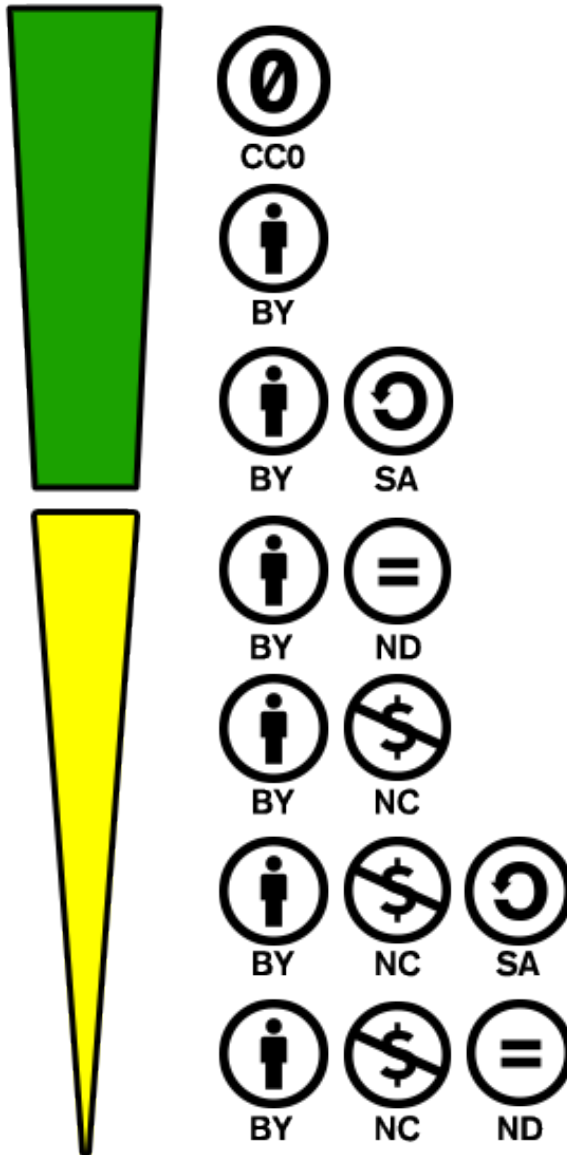
**attribution (BY)** using the data/content requires to give proper credit to the author of the original data/content,

**share-alike (SA)** derivative works require using the same license as their original,

**no-derivative (ND)** forbids making derivative works,

**non-commercial (NC)** forces non-commercial derivation/redistribution.

**MOST OPEN**



**LEAST OPEN**

(from <http://creativecommons.org/examples>)

## Creative Commons Licenses

**Creative Commons CCZero (CC0)** license<sup>2</sup> enforces neither attribution, nor share-alike.

- e.g. Europeana, <http://datahub.io/dataset/europeana-sparql>

**Creative Commons Attribution (CC-BY-4.0)** license<sup>3</sup> enforces attribution, but not share-alike.

- e.g. PLOS<sup>4</sup>, <http://datahub.io/dataset/plos>

**Creative Commons Attribution (CC-BY-SA-4.0)** license<sup>5</sup> enforces attribution, as well as share-alike.

- e.g. DBPedia<sup>6</sup>, <http://dbpedia.org>

---

<sup>2</sup><http://creativecommons.org/publicdomain/zero/1.0/legalcode>

<sup>3</sup><http://creativecommons.org/licenses/by/4.0/>

<sup>4</sup>uses an older version of CC-BY

<sup>5</sup><http://creativecommons.org/licenses/by-sa/4.0/>

<sup>6</sup>uses an older version of CC-BY-SA