Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

Data Interoperability and Semantics
Outline

- < Part 3. Data schemas and semantics >
  - Part 3.1. Data schemas
    - Part 3.1.1. XML Schema
    - Part 3.1.2. JSON Schema
  - Part 3.2. Semantics
    - Part 3.2.1. Heterogeneities and data conflicts
    - Part 3.2.2. Controlled vocabularies and ontologies
    - Part 3.2.3. Resource Description Framework
    - Part 3.2.4. RDFa: Rich structured data markup for web documents
    - Part 3.2.5. JSON-LD: JSON for Linking Data

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.1. Data schemas

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.1. Data schemas

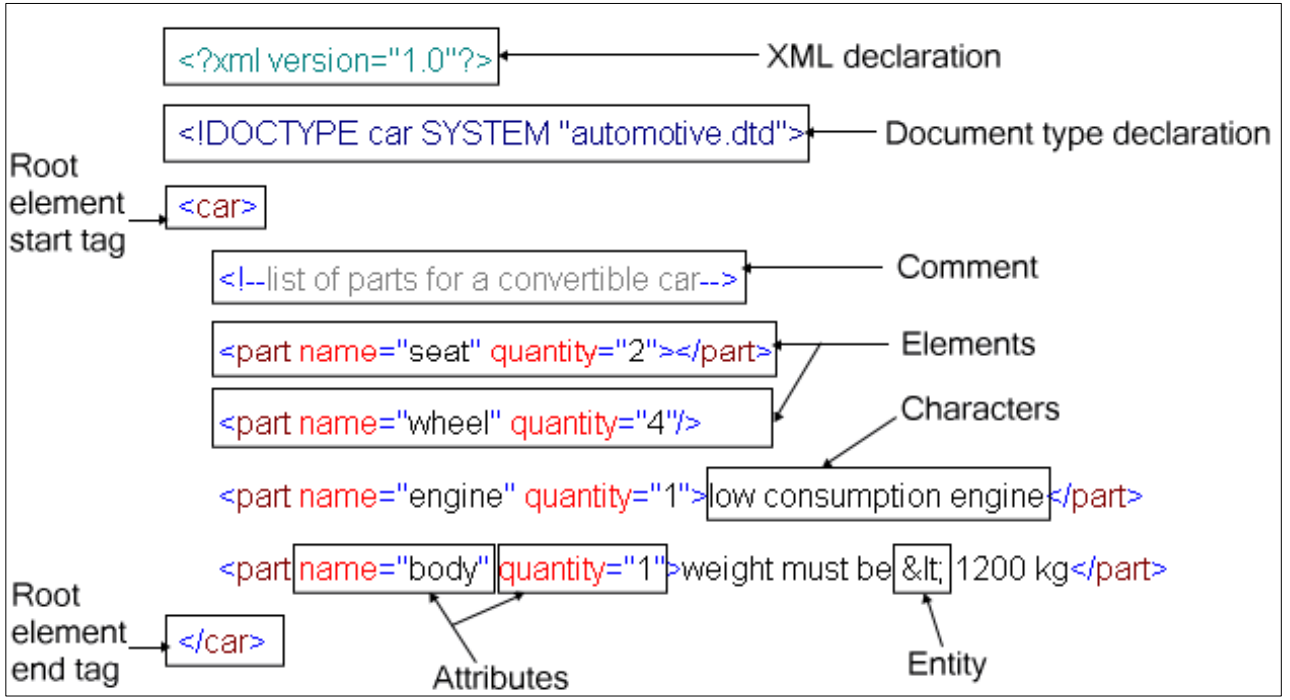
Part 3.1.1. XML Schema

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

Extensible Markup Language

XML (file format)

Filename extension	`.xml`
Internet media type	`application/xml` `text/xml` [1]
Uniform Type Identifier (UTI)	public.xml
UTI conformation	public.text
Magic number	`<?xml`
Developed by	World Wide Web Consortium
Type of format	Markup language
Extended from	SGML
Extended to	Numerous languages, including XHTML · RSS · Atom · KML
Standard	1.0 (Fifth Edition) (November 26, 2008; 12 years ago) 1.1 (Second Edition) (August 16, 2006; 15 years ago)
Open format?	Yes



- v1.0 in 1998, still extensively used in many verticals
- numerous formats based on XML (418 registered on IANA)
  https://en.wikipedia.org/wiki/List_of_XML_markup_languages
  application/atom+xml application/rdf+xml ...
- verbosity, complexity and redundancy

Extensible Markup Language

XML (file format)

Filename extension	`.xml`
Internet media type	`application/xml` `text/xml` [1]
Uniform Type Identifier (UTI)	public.xml
UTI conformation	public.text
Magic number	`<?xml`
Developed by	World Wide Web Consortium
Type of format	Markup language
Extended from	SGML
Extended to	Numerous languages, including XHTML · RSS · Atom · KML
Standard	1.0 (Fifth Edition) (November 26, 2008; 12 years ago) 1.1 (Second Edition) (August 16, 2006; 15 years ago)
Open format?	Yes

Characters and escaping

- unicode implementations: `<?xml version="1.0" encoding="UTF-8"?>`
- escaping characters: `&lt;` '<' - `&amp;` '&' - `&#x2764;` '❤' - etc.

Syntactical correctness

- *well formed vs ill-formed*
- one root tag
- correct nesting
- tag names (approx) start with letter, then alphanumeric or ':'

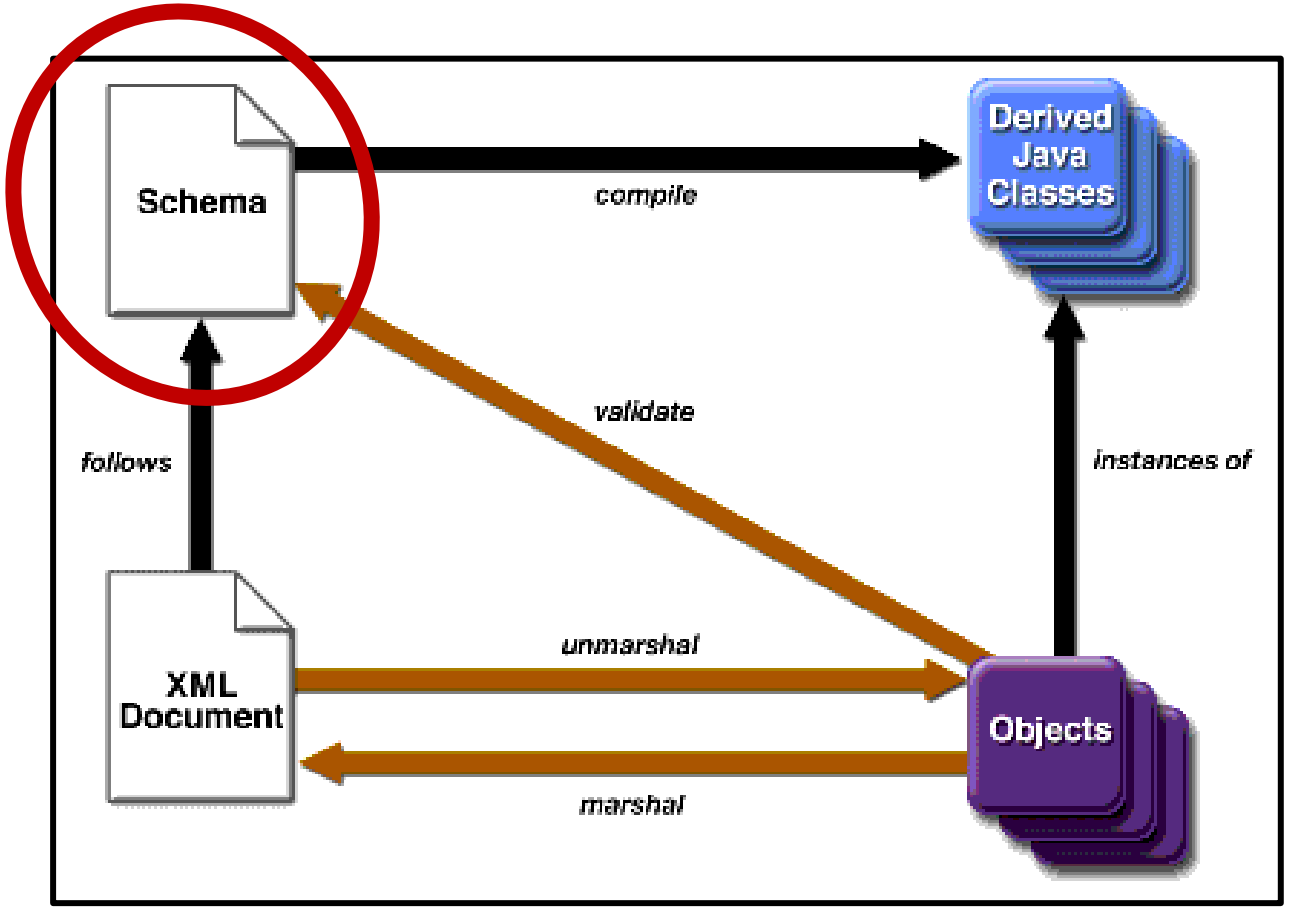
Schemas and validation

- *valid vs invalid*
- DTD, or XML Schema

Namespaces

- `xmlns:ns1="http://example.org/ns1"`
  `xmlns:ns2="http://example.org/ns2"`
- allows to use different schemas together: `<ns1:Tag> <ns2:Tag>`

XML – OOP data binding



XML-OOP data binding

ex Java: https://zetcode.com/java/jaxb/

Employee.java

```java
@XmlRootElement(name = "employee")
@XmlAccessorType(XmlAccessType.FIELD)
public class Employee implements Serializable
{
    private Integer id;
    private String firstName;
    private String lastName;
}
```

employee.xml

```xml
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<employee>
    <firstName>Lokesh</firstName>
    <id>1</id>
    <lastName>Gupta</lastName>
</employee>
```

Java Architecture for XML Binding (JAXB) example

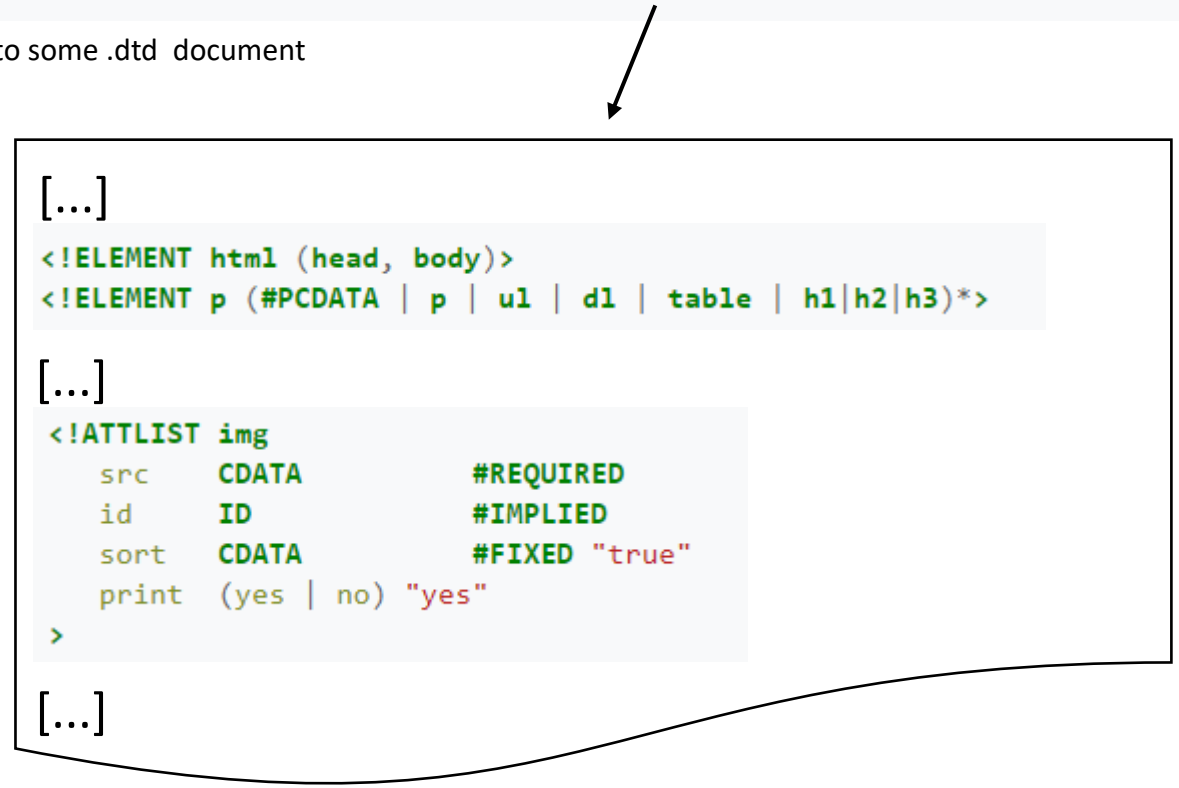
https://howtodoinjava.com/jaxb/jaxb-annotations/

XML Schema

- Document Type Definition (DTD, 2008)

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
```

link to some .dtd document



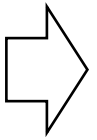
[...]

```
<!ELEMENT html (head, body)>
<!ELEMENT p (#PCDATA | p | ul | dl | table | h1|h2|h3)*>
```

[...]

```
<!ATTLIST img
   src     CDATA          #REQUIRED
   id      ID             #IMPLIED
   sort    CDATA          #FIXED "true"
   print   (yes | no) "yes"
>
```

[...]



```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html>
<!-- the XHTML document body starts here-->
<html xmlns="http://www.w3.org/1999/xhtml">
 ...
</html>
```

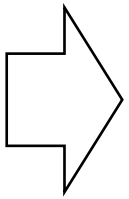
HTML 5 : no more link to a DTD

XML Schema

- Regular Language for XML Next Generation (RELAX NG, 2001) https://relaxng.org/

```
<element name="addressBook" xmlns="http://relaxng.org/ns/structure/1.0">
  <zeroOrMore>
    <element name="card">
      <element name="name">
        <text/>
      </element>
      <element name="email">
        <text/>
      </element>
    </element>
  </zeroOrMore>
</element>
```

RELAX NG document



```
<addressBook>
  <card>
    <name>John Smith</name>
    <email>js@example.com</email>
  </card>
  <card>
    <name>Fred Bloggs</name>
    <email>fb@example.net</email>
  </card>
</addressBook>
```

Valid XML document

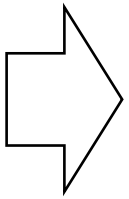
XML Schema

- XML Schema (W3C, 2003)

  The only one you should use (if you need to)

```xml
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">

<xs:element name="note">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="to" type="xs:string"/>
      <xs:element name="from" type="xs:string"/>
      <xs:element name="heading" type="xs:string"/>
      <xs:element name="body" type="xs:string"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>

</xs:schema>
```

XML Schema document



```xml
<?xml version="1.0"?>

<note
xmlns="https://www.w3schools.com"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="https://www.w3schools.com/xml note.xsd">
  <to>Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
```

Valid XML document

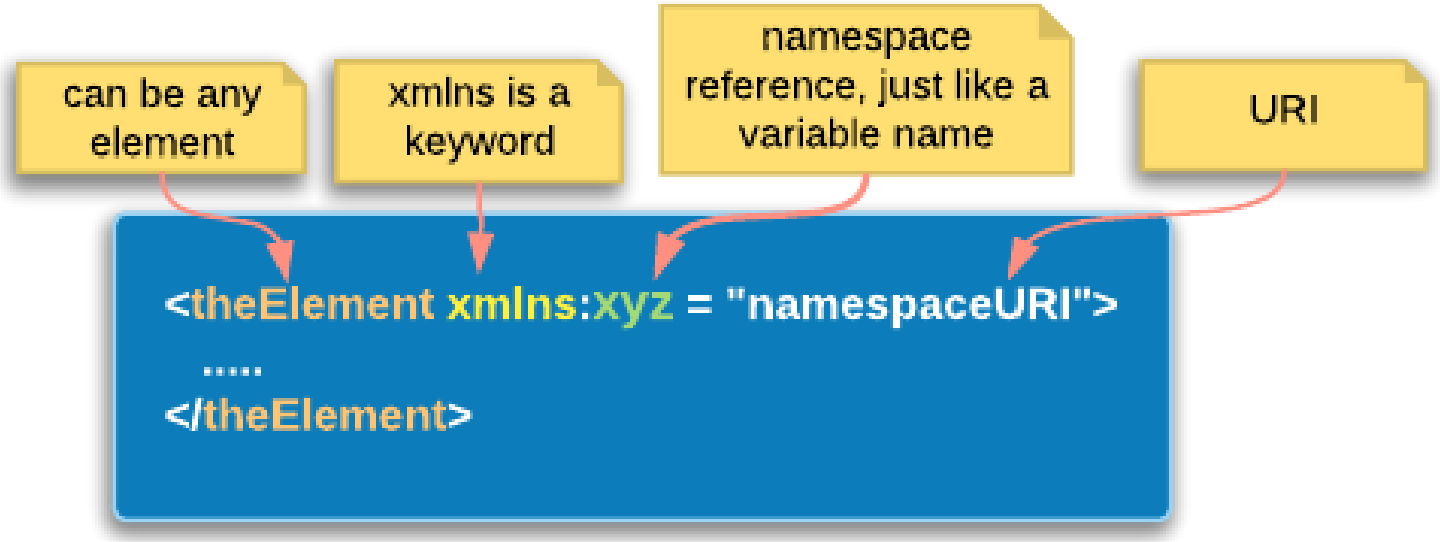
https://www.w3schools.com/xml/schema_intro.asp

XML Schema

- Tutorials
  - https://www.tutorialspoint.com/xml/xml_schemas.htm
  - https://www.w3schools.com/xml/schema_intro.asp
  - …
- What to do with XML Schemas
  - validate documents
  - generate forms
  - generate classes (any OOP language)
  - …

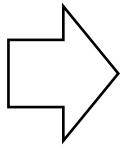
Multiple XML Schemas ?

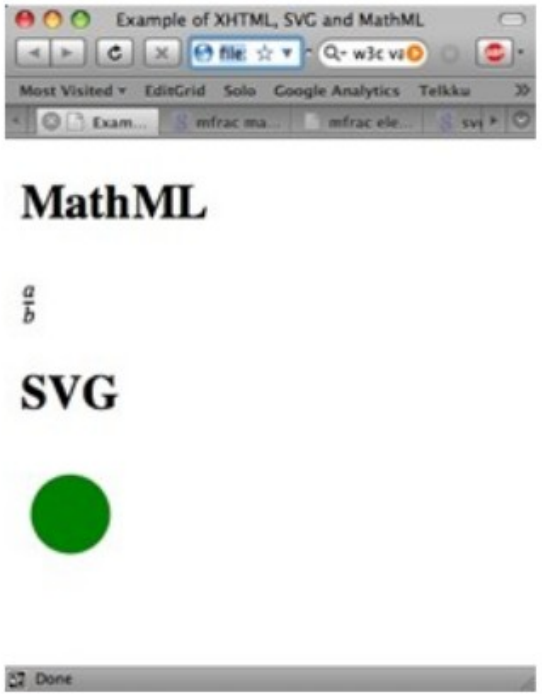
- How to combine different XML schemas in a single file?
- Using XML namespaces



Example: A document with XHTML + MathML + SVG

```
<?xml version="1.0"?>
<!DOCTYPE html PUBLIC
   "-//W3C//DTD XHTML 1.1 plus MathML 2.0 plus SVG 1.1//EN"
   "http://www.w3.org/2002/04/xhtml-math-svg/xhtml-math-svg-flat.dtd">
<html xmlns      = "http://www.w3.org/1999/xhtml"
    xmlns:svg    = "http://www.w3.org/2000/svg">

 <head>
  <title>Example of XHTML, SVG and MathML</title>
 </head>
 <body>

  <h2>MathML</h2>
  <p>
   <math xmlns="http://www.w3.org/1998/Math/MathML">
    <mfrac>
     <mi>a</mi>
     <mi>b</mi>
    </mfrac>
   </math>
  </p>

  <h2>SVG</h2>

  <p>
   <svg:svg width="50px" height="50px">
    <svg:circle cx="25px" cy="25px" r="20px" fill="green"/>
   </svg:svg>
  </p>

 </body>
</html>
```





Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.1. Data schemas

Part 3.1.2. JSON Schema

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

JavaScript Object Notation

JavaScript Object Notation

Filename extension	`.json`
Internet media type	`application/json`
Type code	TEXT
Uniform Type Identifier (UTI)	public.json
Type of format	Data interchange
Extended from	JavaScript
Standard	STD 90 (RFC 8259), ECMA-404, ISO/IEC 21778:2017
Open format?	Yes
Website	json.org

JSON Schema

JSON Schema Validator

newtonsoft.com

An online, interactive JSON Schema validator. Supports JSON Schema Draft 3, Draft 4, Draft 6 and Draft 7.

View source code

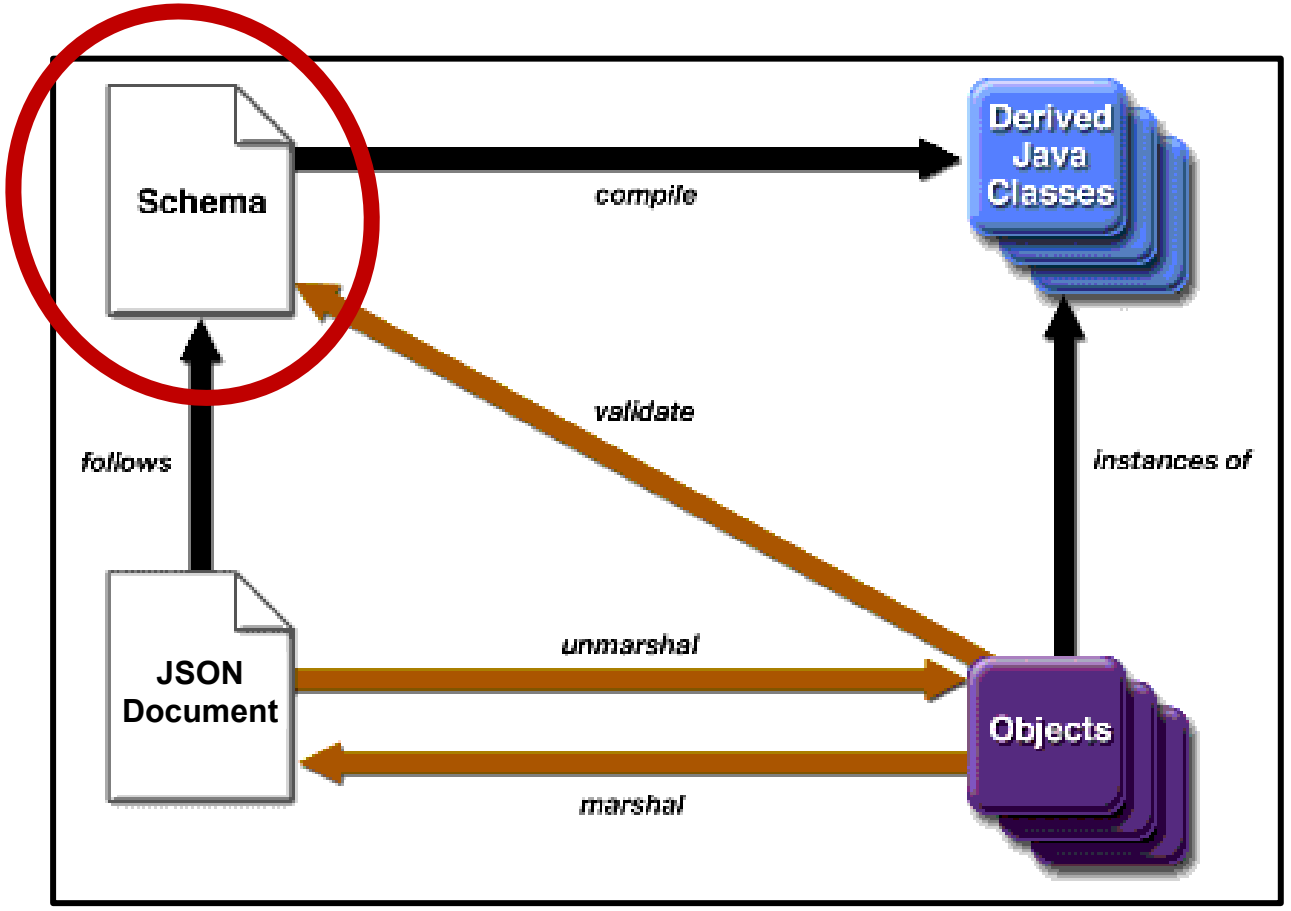
Select schema: Empty schema

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",

  "definitions": {
    "person": {
      "type": "object",
      "properties": {
        "name": { "type": "string" },
        "children": {
          "type": "array",
          "items": { "$ref": "#/definitions/person" },
          "default": []
        }
      }
    },

  "type": "object",

  "properties": {
    "person": { "$ref": "#/definitions/person" }
  }
}
```

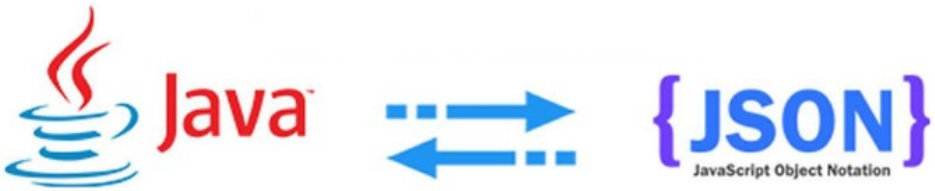
Input JSON:

```
{
  "personsdfsd": {
    "nameklk": "Elizabeth",
    "childrensss": [
      {
        "namess": "Charles",
        "childrenss": [
          {
            "namess": "William",
            "children": [
              { "namesss": "George" },
              { "name": "Charlotte" }
            ]
          },
          {
            "namesss": "Harry"
          }
        ]
      }
    ]
  }
}
```

JSON – OOP data binding



JSON-OOP data binding



JSON Schema Specification (latest)

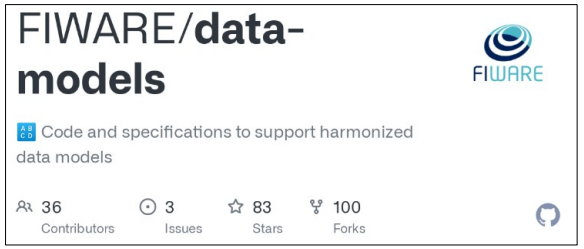
https://json-schema.org/draft/2020-12/json-schema-validation.html

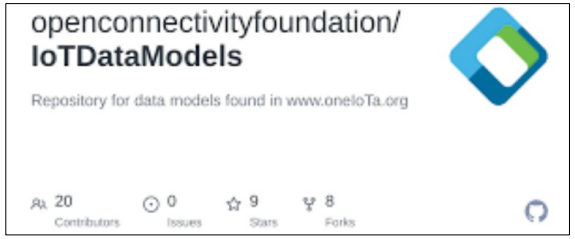
Table of Contents

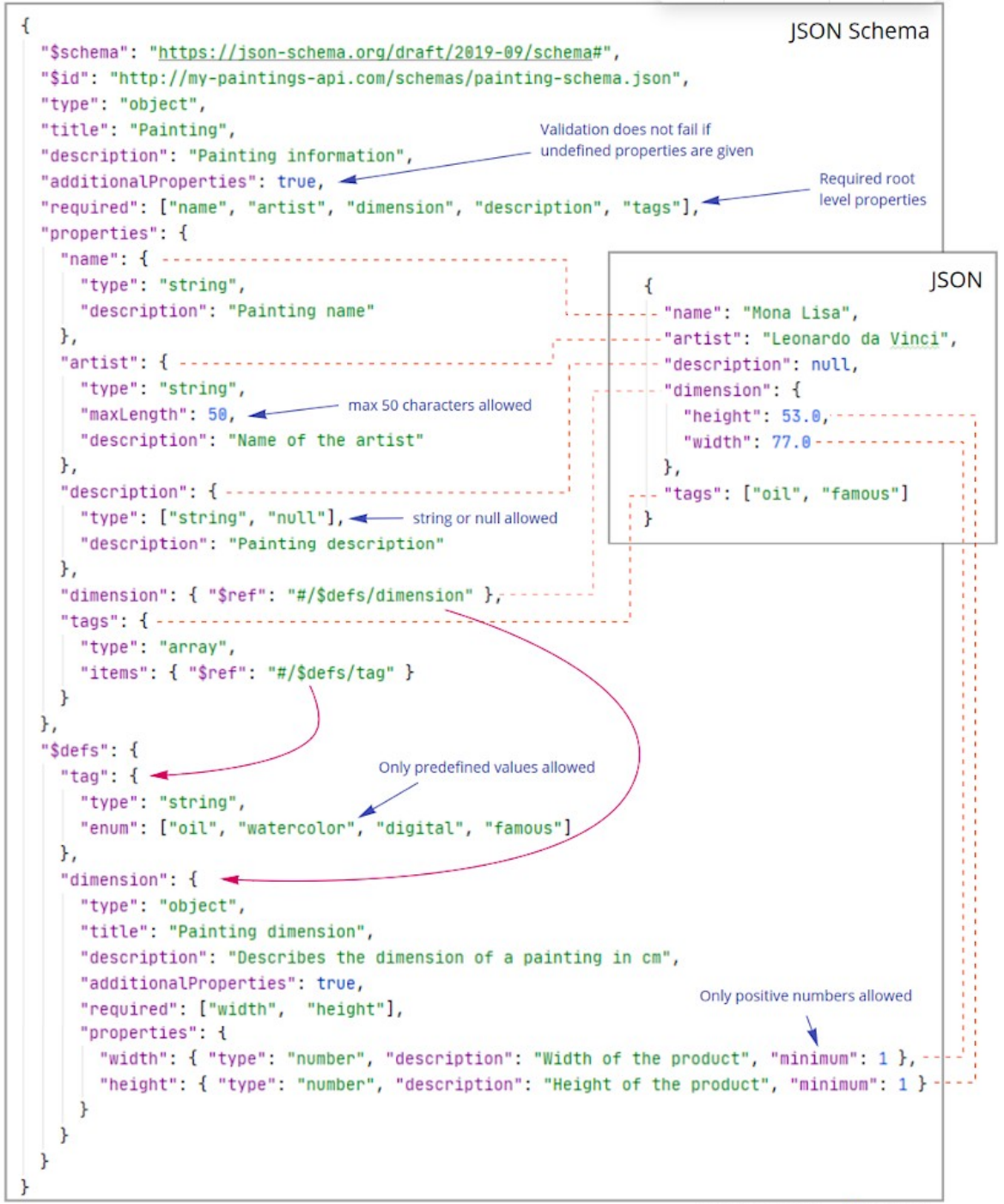
JSON Schema

Examples:

- https://json-schema.org/learn/miscellaneous-examples.html
- https://www.fiware.org/developers/data-models/
- https://oneiota.org/







Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.2. Semantics

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.2. Semantics

Part 3.2.1. Heterogeneities and data conflicts

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

DEMO: Compare JSON documents

https://samples.openweathermap.org/data/2.5/weather?id=2172797&appid=b6907d289e10d714a6e88b30761fae22

and

https://www.prevision-meteo.ch/services/json/lausanne

and

https://github.com/smart-data-models/dataModel.Weather/blob/master/WeatherForecast/examples/example-normalized.json

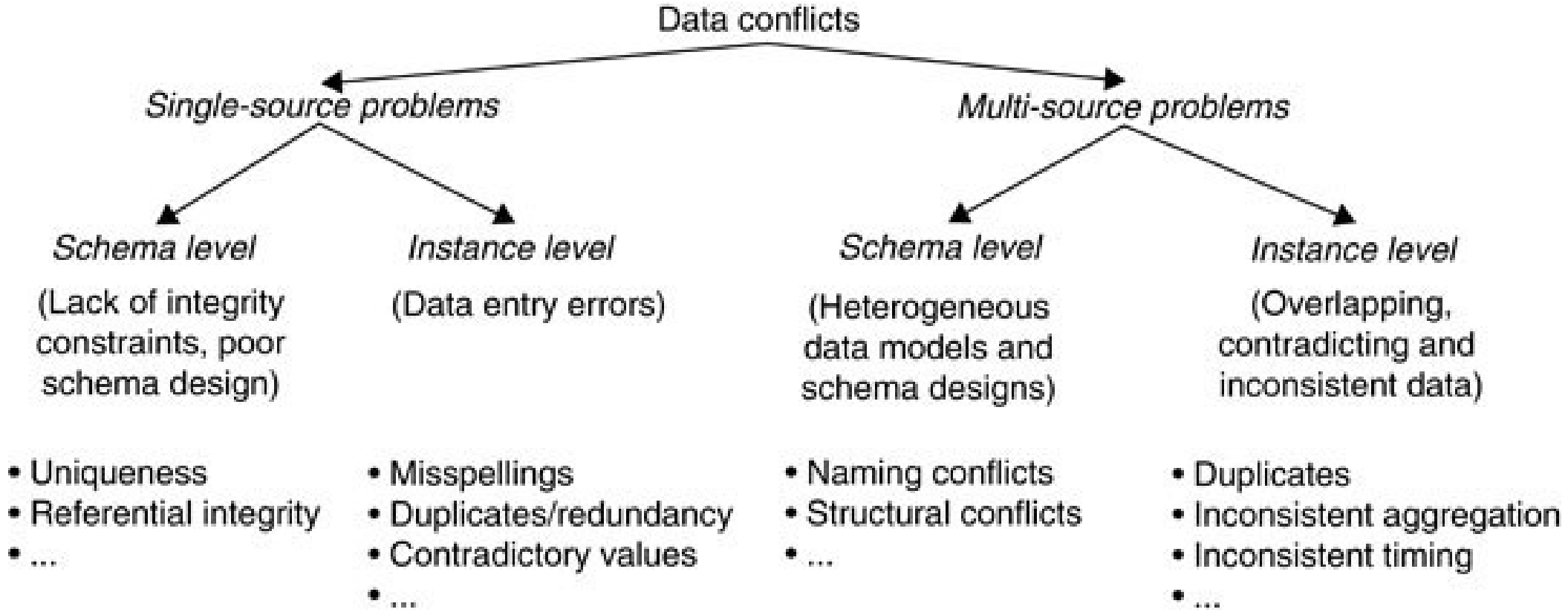
DEMO: Compare JSON documents

Different modeling choices were made, which make these two services completely non-interoperable:

- the lat/long coordinates: string vs number
- the UNIX timestamps vs dates and times
- the choice of keys and the semantics (meaning) of the values
- the units of temperature, pressure, wind speed, …
- the semantics of wind direction
- the value for "icon": "03n" - (if we follow our nose on the website, we may figure out it refers to http://openweathermap.org/img/w/03n.png)
- the country codes ISO 3166-1 ALPHA-2 and ISO 3166-1 ALPHA-3 (example of Australia and Austria)

Data conflicts

*Data conflicts are **deviations between data intended to capture the same state of a real-world entity**. Data with conflicts are often called "dirty" data and can mislead analysis performed on it.*



Do HH. (2009) Data Conflicts. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_97

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.2. Semantics

Part 3.2.2. Controlled vocabularies, thesauri, taxonomies

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

Controlled vocabularies

… an established list of standardized terminology for use in indexing and retrieval of information

— OECD

… an organized arrangement of words and phrases used to index content and/or retrieve content through browsing or searching

— Getty institute

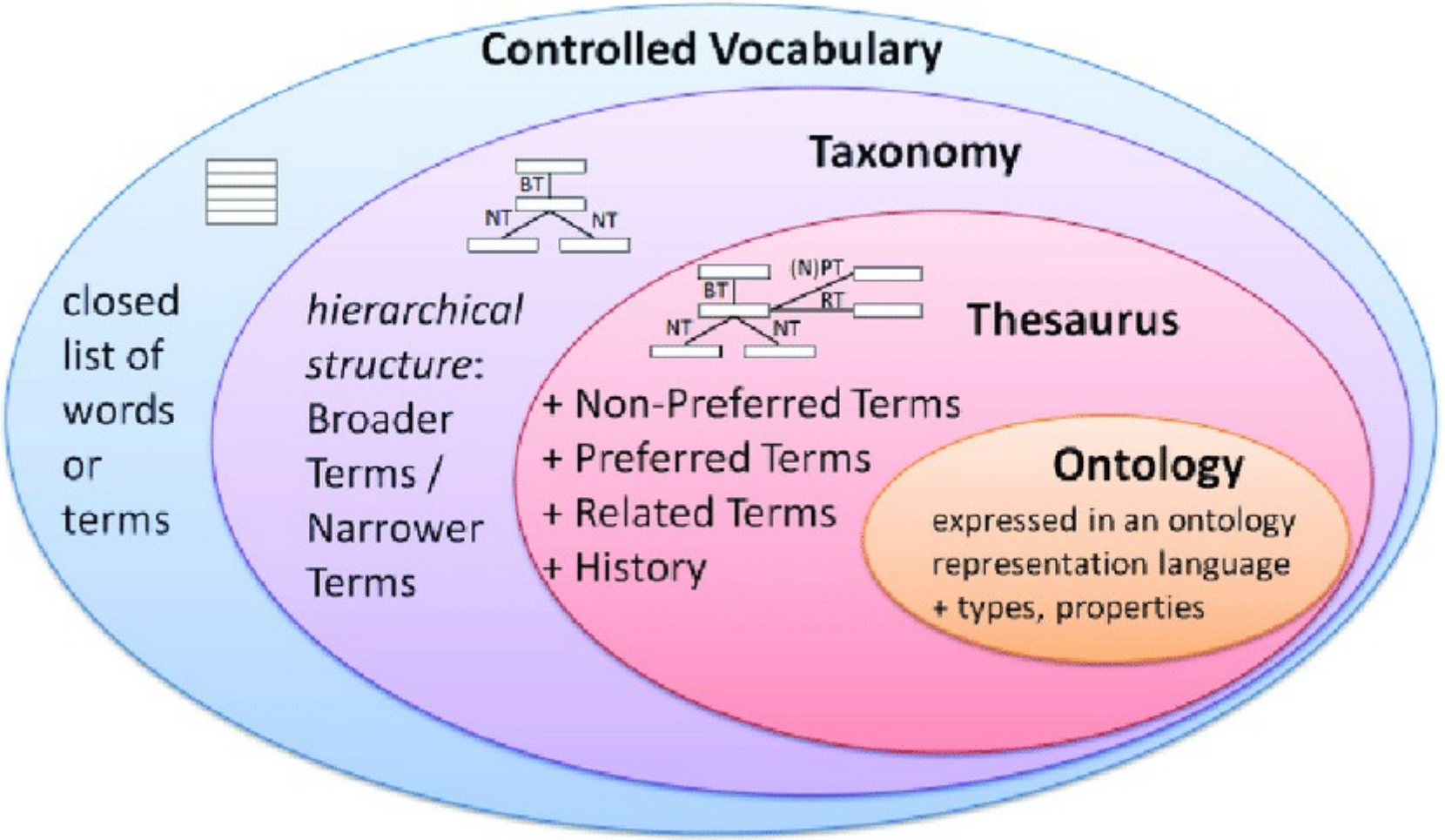
… an standardized – yet dynamic – set of terms and phrases authorized for use in an indexing system to describe a subject area or information domain

— SCIP

vocabulary for which the entries, i.e. **definition**/**term** pairs, are controlled by a **Source Authority** based on a **rulebase** and **process** for addition/deletion of entries

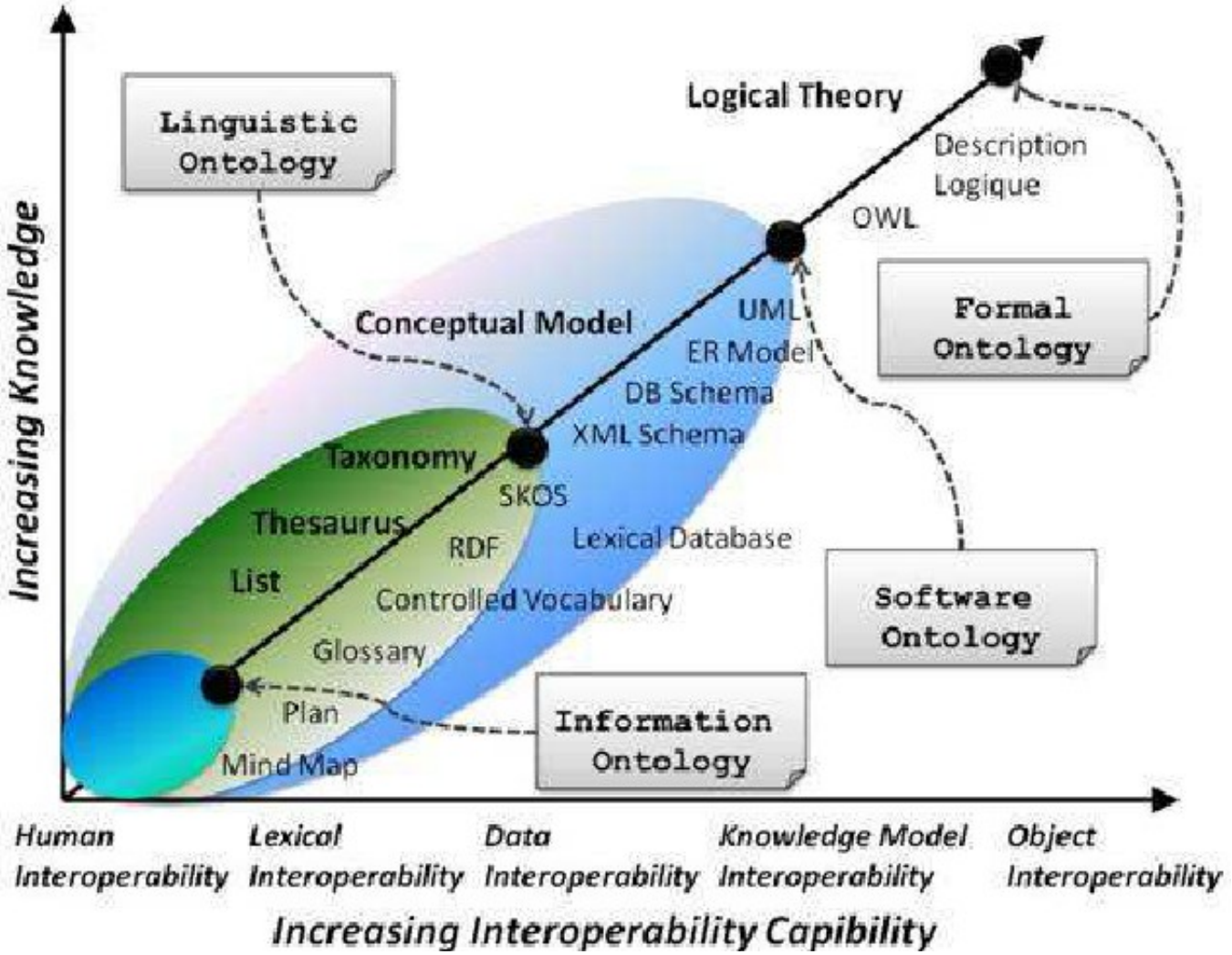
— ISO/IEC 15944-5:2008(en)
Information technology — Business Operational View

Controlled vocabulary, taxonomy, thesaurus, ontology



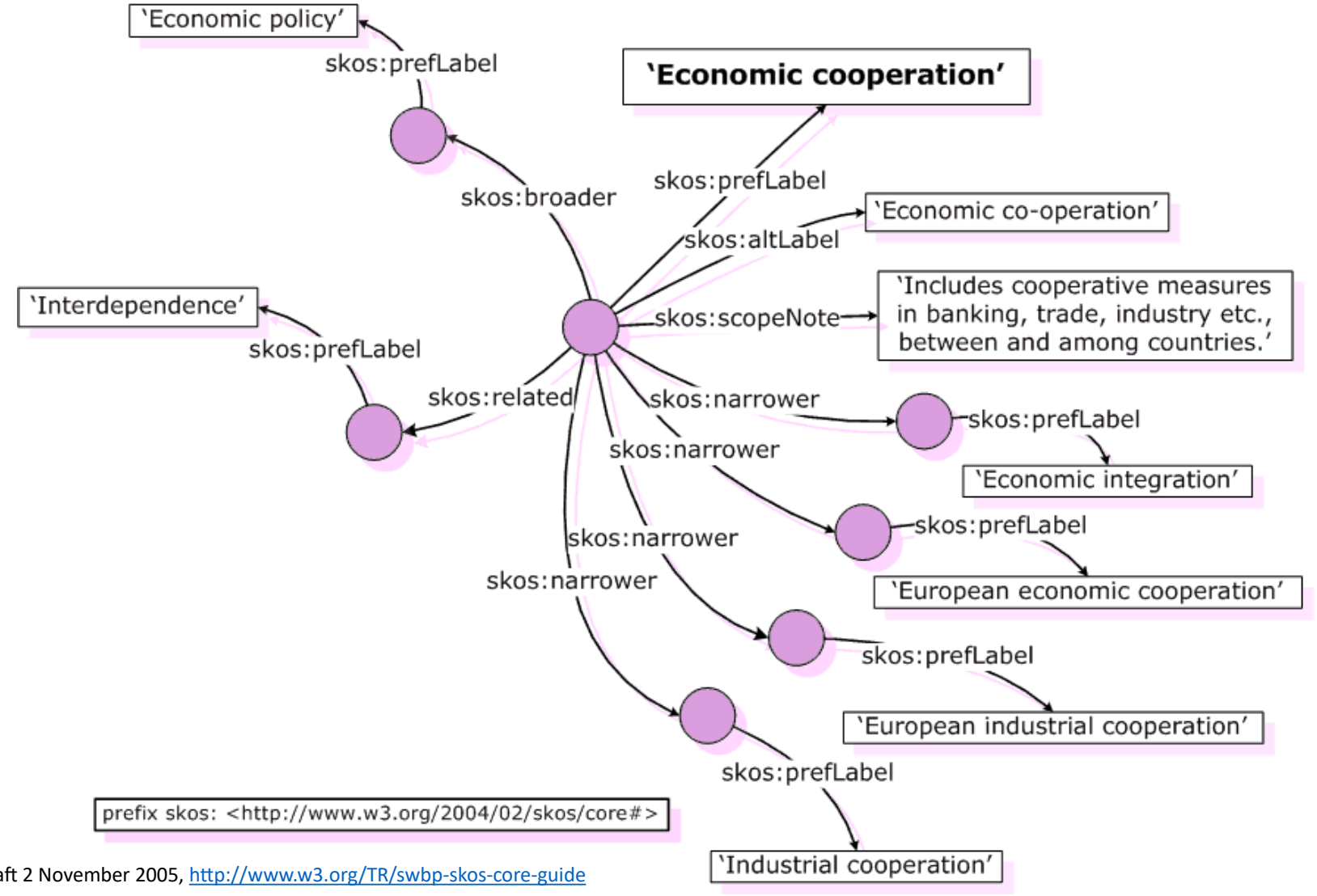
Kopácsi, Sándor & Hudak, Rastislav & Ganguly, Raman. (2017). Implementation of a Classification Server to Support Metadata Organization for Long Term Preservation Systems. Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare. 70. 225. 10.31263/voebm.v70i2.1897.

Controlled vocabulary, taxonomy, thesaurus, ontology



Roussey, Catherine & Pinet, François & Kang, Myoung-Ah & Corcho, Oscar & Falquet, Gilles & Métral, Claudine & Teller, Jacques & Tweed, Christopher. (2011). Ontologies for Interoperability. 10.1007/978-0-85729-724-2_3.

SKOS Simple Knowledge Organization System



SKOS Core Guide, W3C Working Draft 2 November 2005, http://www.w3.org/TR/swbp-skos-core-guide

Examples of thesauri, taxonomies, …

UNESCO Thesaurus

The UNESCO Thesaurus is **a controlled and structured list of terms used in subject analysis and retrieval of documents and publications** in the fields of education, culture, natural sciences, social and human sciences, communication and information.

https://skos.um.es › unescothes

UNESCO Thesaurus - SKOS

GEMET General Multilingual Environmental Thesaurus

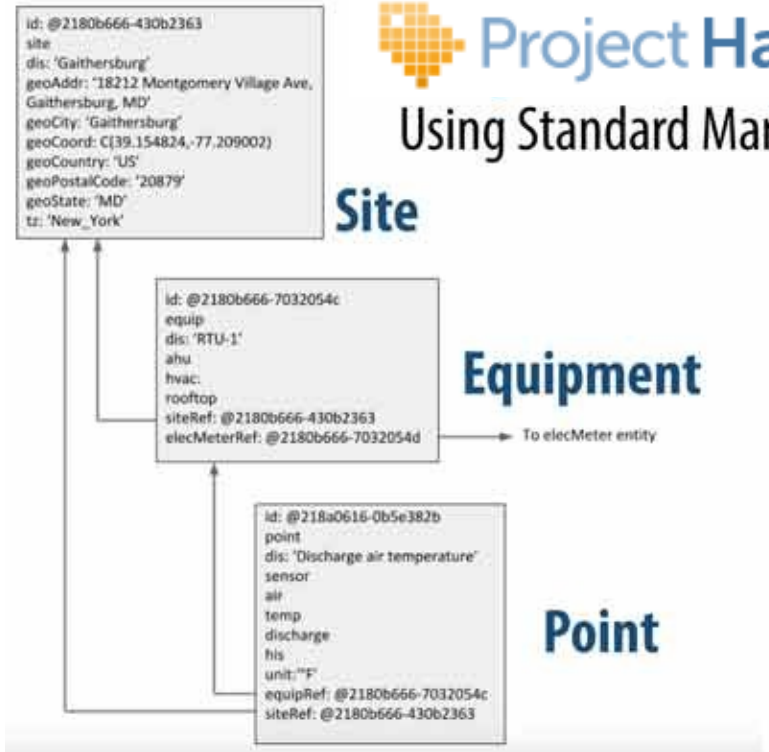
https://www.eionet.europa.eu/gemet/en/webservices/

nasa/**dictionaries**

A collection of NASA "dictionaries", including thesauri, taxonomies and ontologies.

2 Contributors · 2 Issues · 51 Stars · 31 Forks

https://github.com/nasa/dictionaries



Project Haystack

Using Standard Markup for…

https://project-haystack.org/

List at https://www.w3.org/2001/sw/wiki/SKOS/Datasets

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.2. Semantics

Part 3.2.3. Resource Description Framework

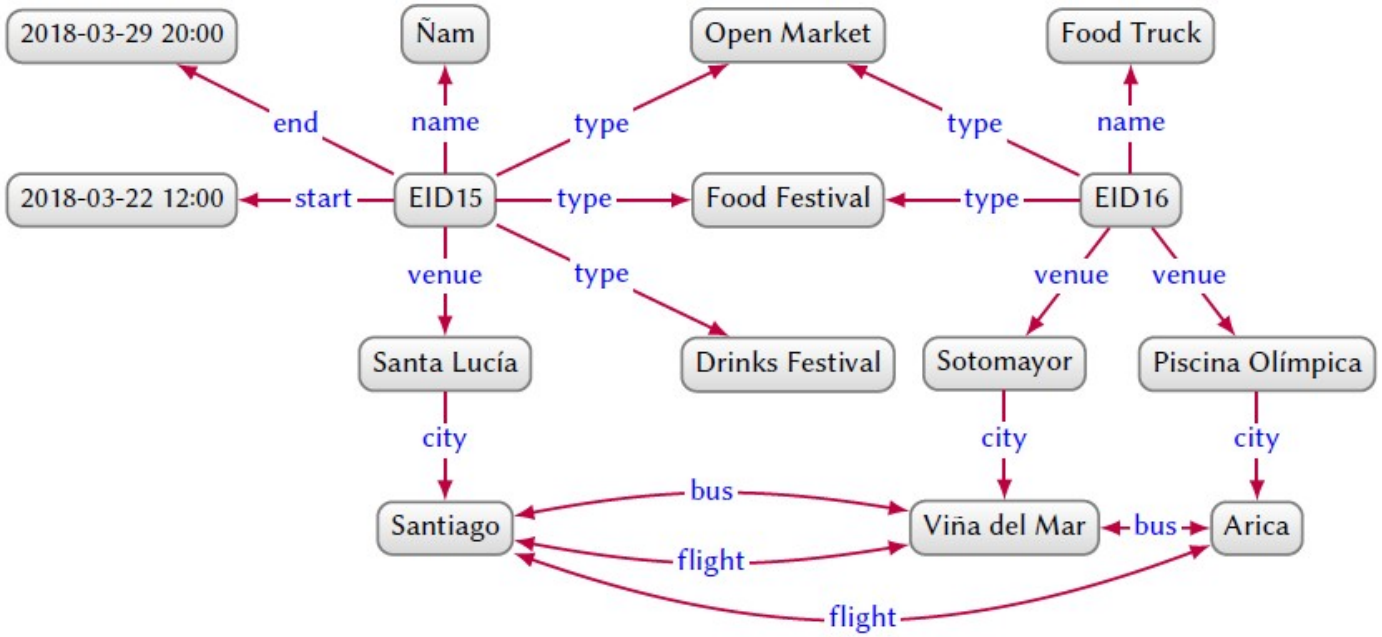
ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

RDF- **R**esource **D**escription **F**ramework

RDF

Resource Description Framework

Abbreviation	RDF
Status	Published, W3C Recommendation
Year started	1996; 25 years ago
First published	February 28, 1996; 25 years ago
Latest version	1.1 (Recommendation) February 25, 2014; 7 years ago
Organization	World Wide Web Consortium (W3C)
Committee	RDF Working Group
Editors	Richard Cyganiak, David Wood, Markus Lanthaler
Base standards	URI
Related standards	RDFS, OWL, RIF, RDFa
Domain	Semantic Web
Website	www.w3.org/TR/rdf11-concepts/



An RDF Graph

Vertices:

Resources: Things. Uniquely identified by URIs

Literals: Unicode string + Datatype identified by URI. *out-degree=0*

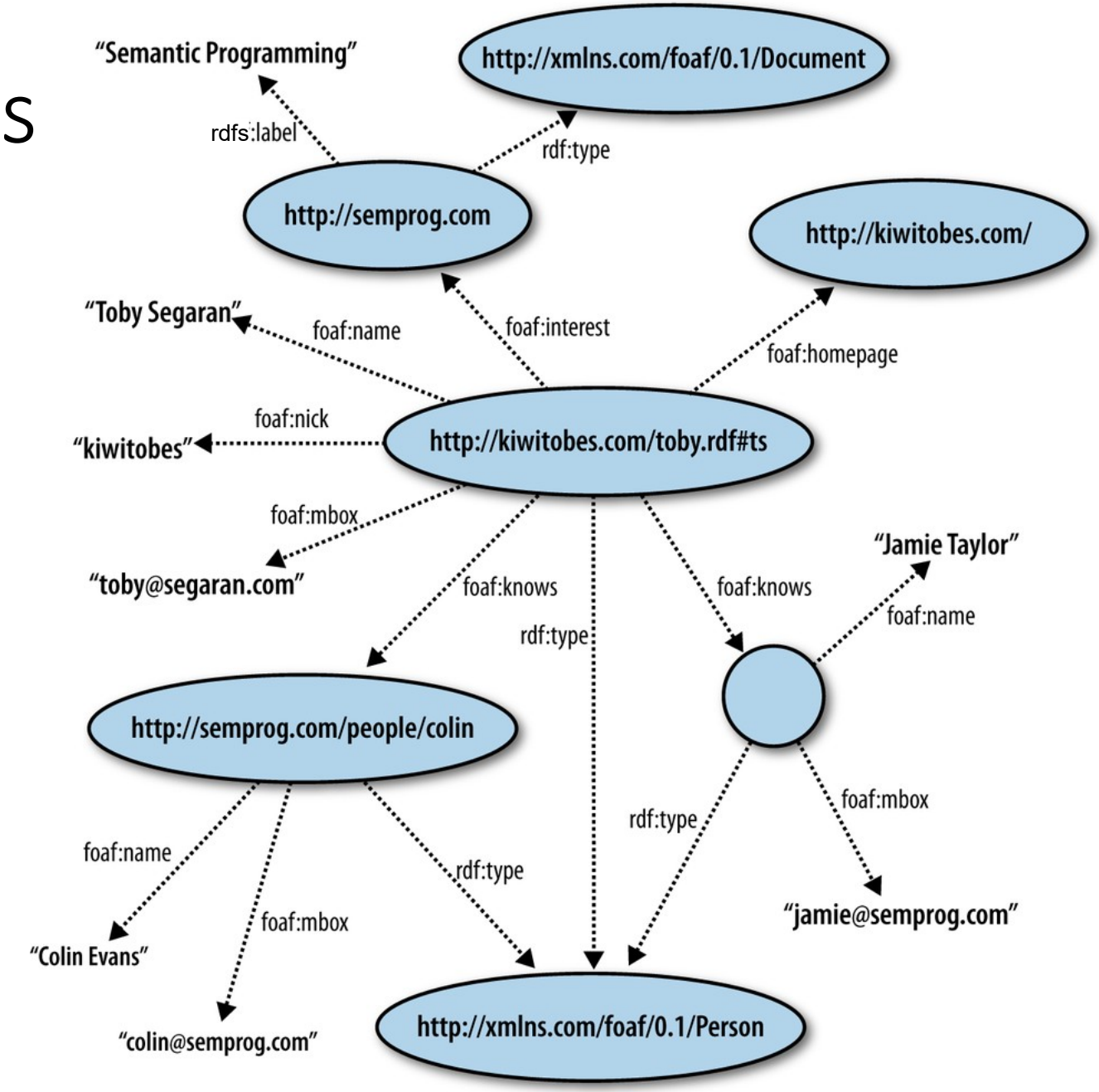
Edges:

Relationships: connections between vertices. Uniquely identified by URIs

W3C Semantic Web

CURIE = Compact URIs

```
@prefix rdf:
   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs:
   <http://www.w3.org/2000/01/rdf-schema#>.
@prefix foaf:
   <http://xmlns.com/foaf/0.1/>.
@prefix schema:
   <http://schema.org/>.
...
```



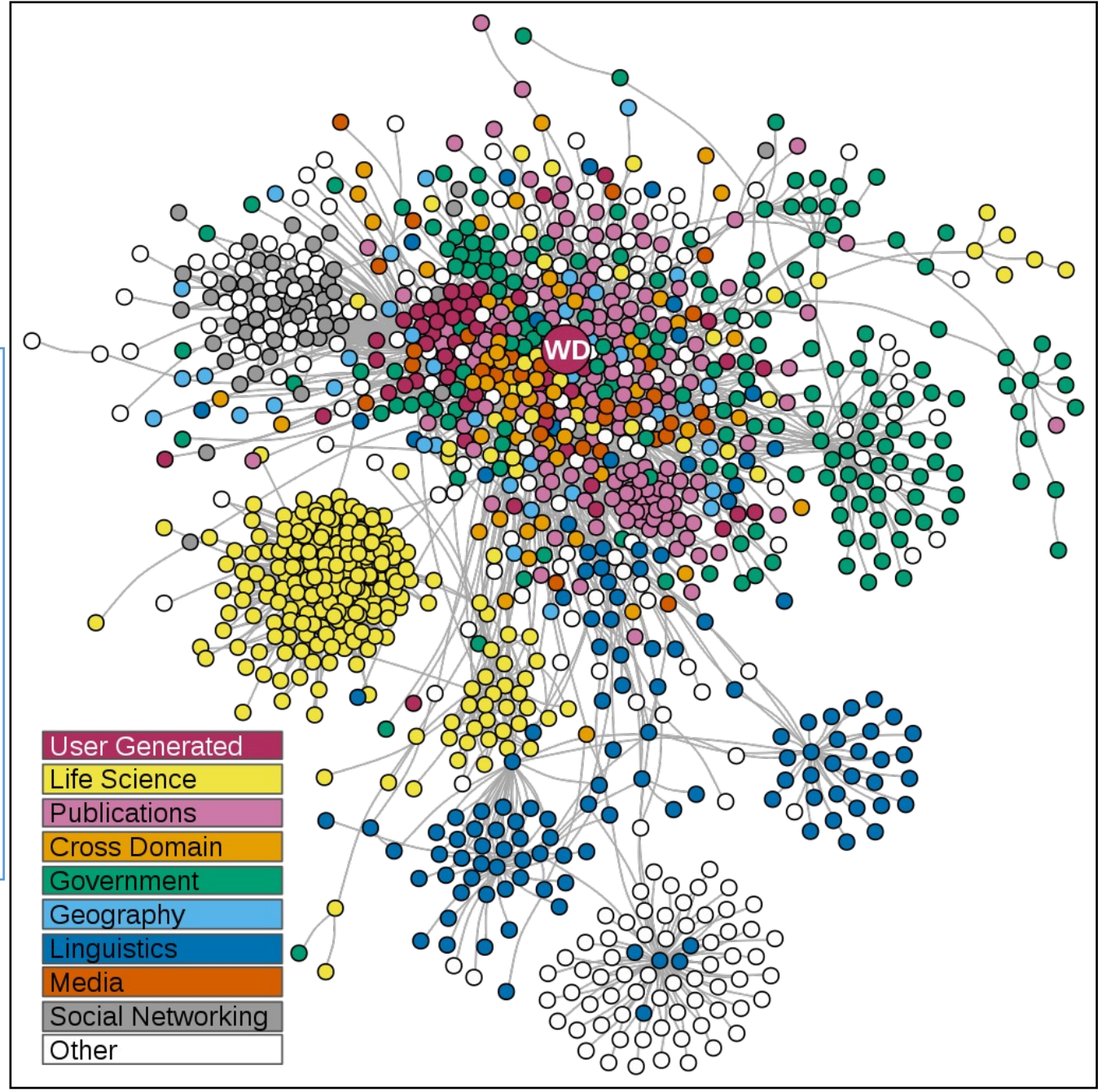
The Linked Data

It's possible to use URIs defined in an external graph

- to use a reference identification system
- to augment a graph

anyone can say anything about anything

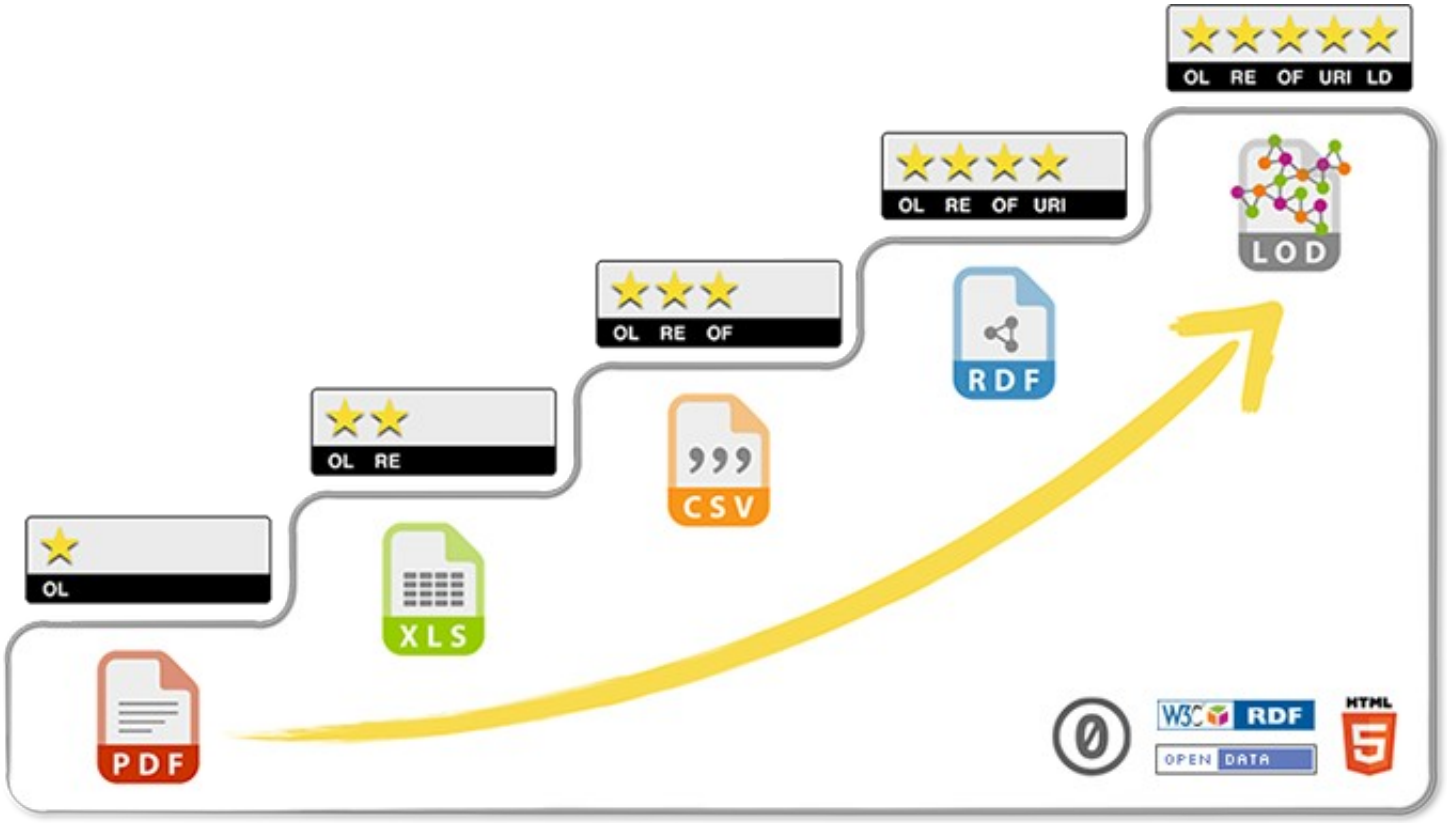
One can specify vertice co-reference with some special edges: owl:sameAs from W3C OWL vocabulary



Linked Open Data cloud visualisation: each node is a RDF dataset. Links indicate the existence of external identification links. Source: wikipedia

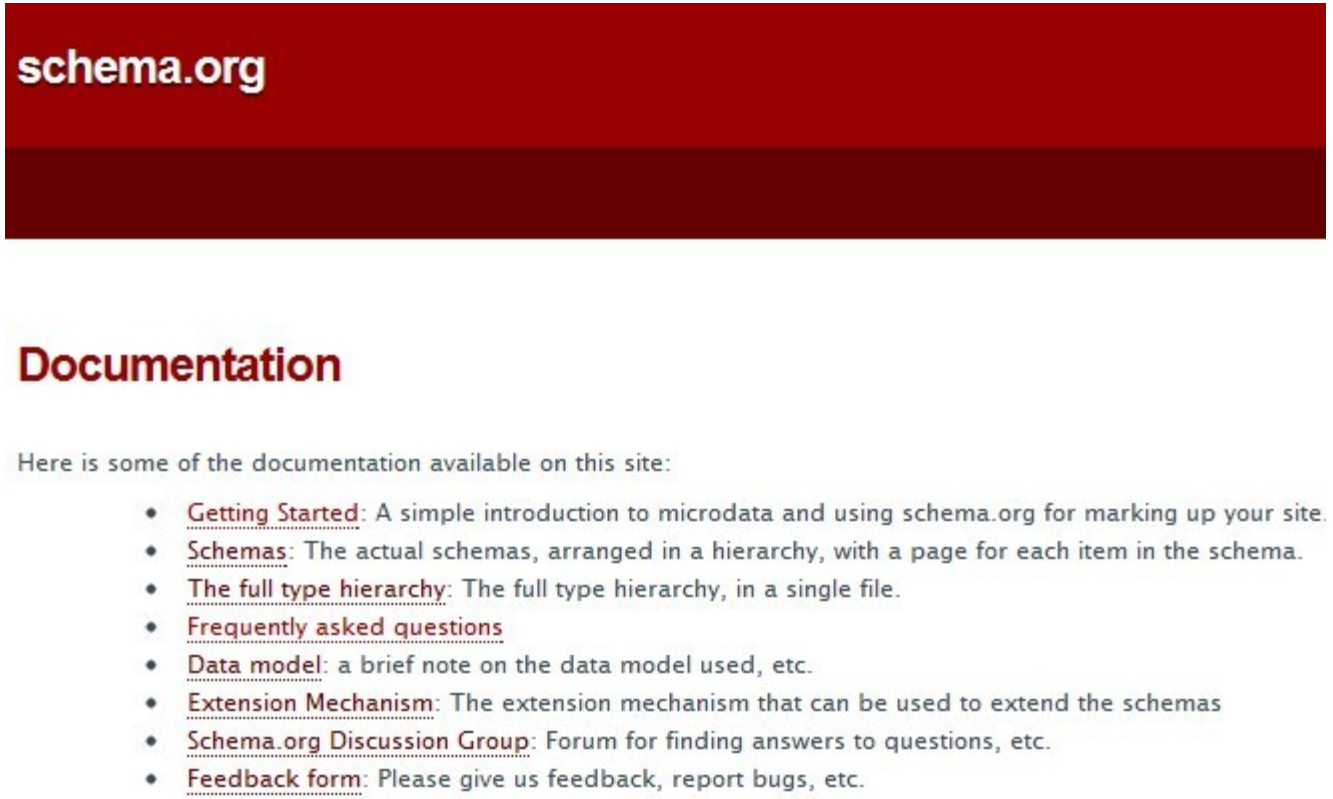
5 ★ OPEN DATA

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data.



https://5stardata.info/en/

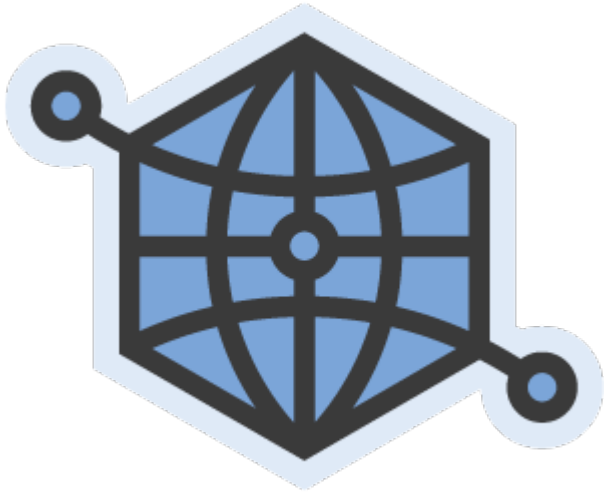
Examples of RDF vocabularies & ontologies



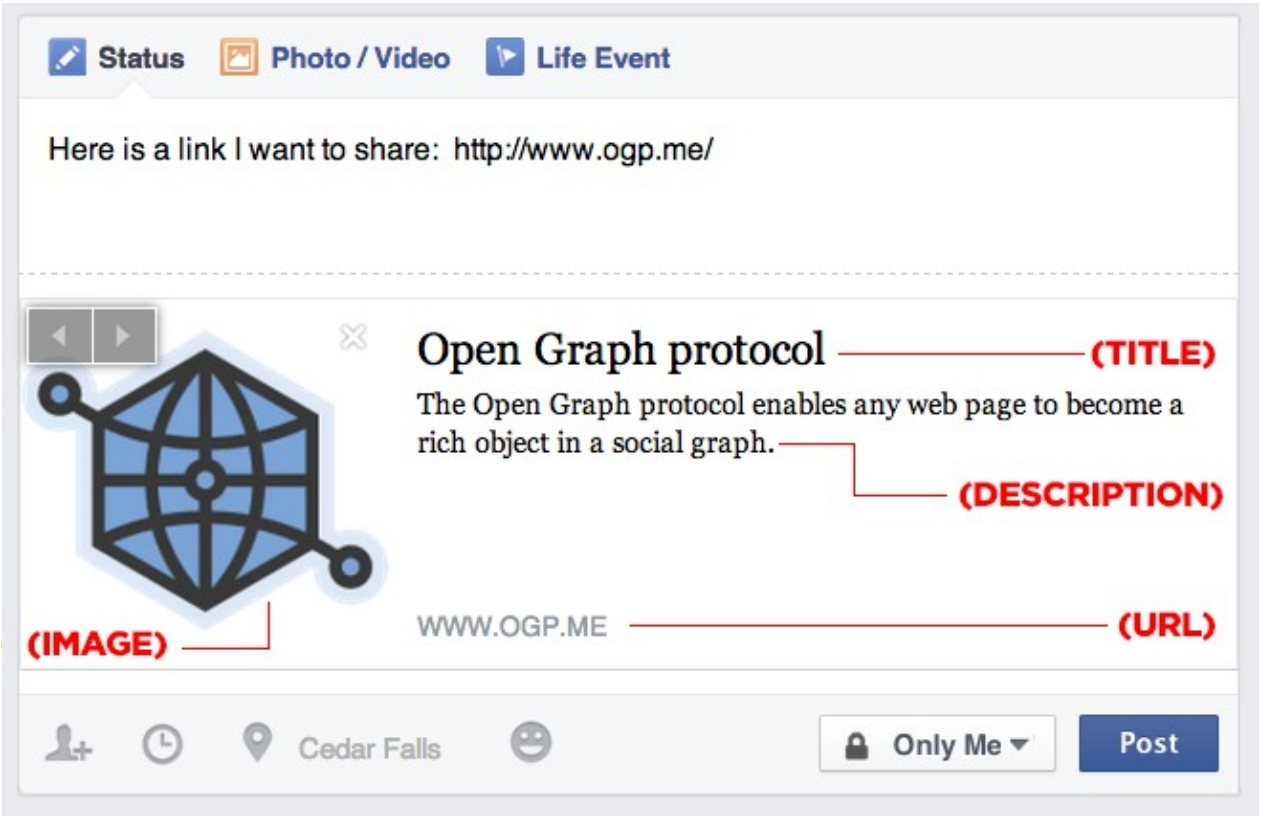
Over 10 million sites use Schema.org to markup their web pages and email messages. Many applications from Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences.

Founded by Google, Microsoft, Yahoo and Yandex, 2011
Test with https://validator.schema.org/ or https://developers.google.com/search/docs/advanced/structured-data

Examples of RDF vocabularies & ontologies



```
<html prefix="og: https://ogp.me/ns#">
<head>
<title>The Rock (1996)</title>
<meta property="og:title" content="The Rock" />
<meta property="og:type" content="video.movie" />
<meta property="og:url" content="https://www.imdb.com/title/tt01175
<meta property="og:image" content="https://ia.media-imdb.com/images
...
</head>
...
</html>
```



Open Graph Protocol https://ogp.me/ (Facebook, 2010), test with https://developers.facebook.com/tools/debug/

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.2. Semantics

Part 3.2.4. RDFa: Rich structured data markup for web documents

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

RDFa Resource Description Framework in Attributes

See RDFa 1.1 Primer - Rich Structured Data Markup for Web Documents - https://www.w3.org/TR/rdfa-primer/

```
<html prefix="og: https://ogp.me/ns#">
<head>
<title>The Rock (1996)</title>
<meta property="og:title" content="The Rock" />
<meta property="og:type" content="video.movie" />
<meta property="og:url" content="https://www.imdb.com/title/tt0117500/" />
<meta property="og:image" content="https://ia.media-imdb.com/images/rock.jpg" />
...
</head>
...
</html>
```

```
<div typeof="foaf:Person" xmlns:foaf="http://xmlns.com/
foaf/0.1/">
   <p property="foaf:name">
     Alice Birpemswick
   </p>
   <p>
     Email: <a rel="foaf:mbox"
href="mailto:alice@example.com">alice@example.com</a>
   </p>
   <p>
     Phone: <a rel="foaf:phone" href="tel:
+1-617-555-7332">+1 617.555.7332</a>
   </p>
</div>
```

RDFa Resource Description Framework in Attributes

See RDFa 1.1 Primer - Rich Structured Data Markup for Web Documents - https://www.w3.org/TR/rdfa-primer/

The Goal

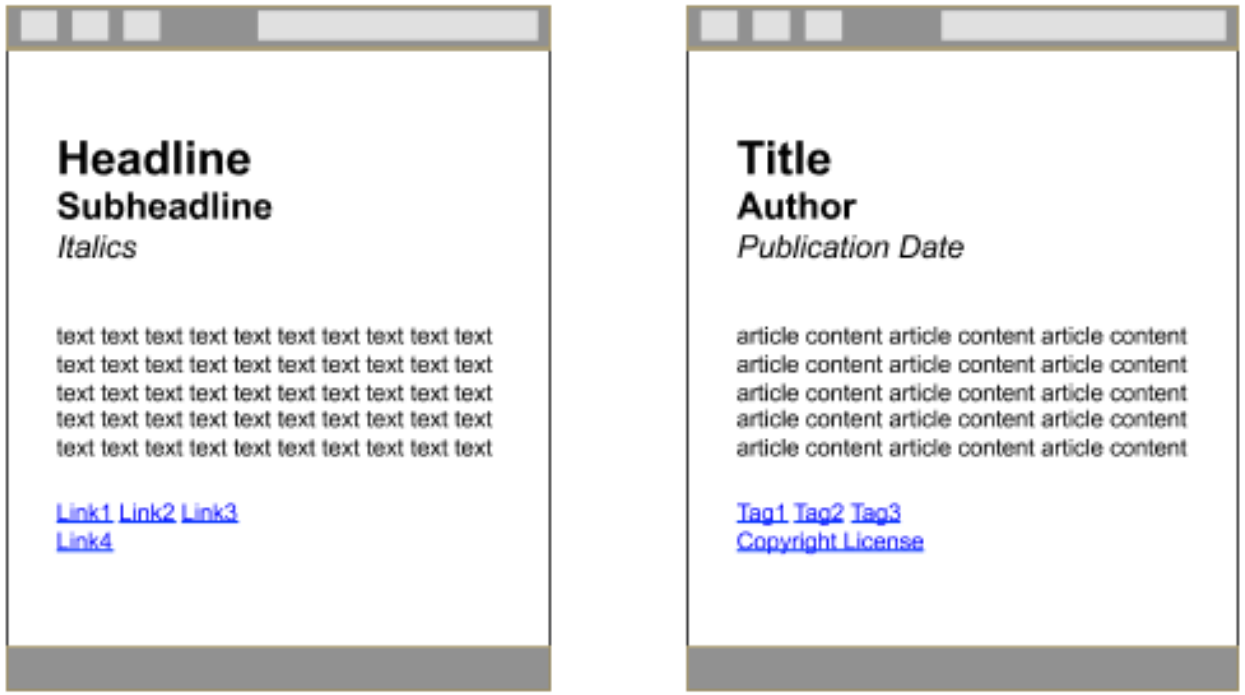
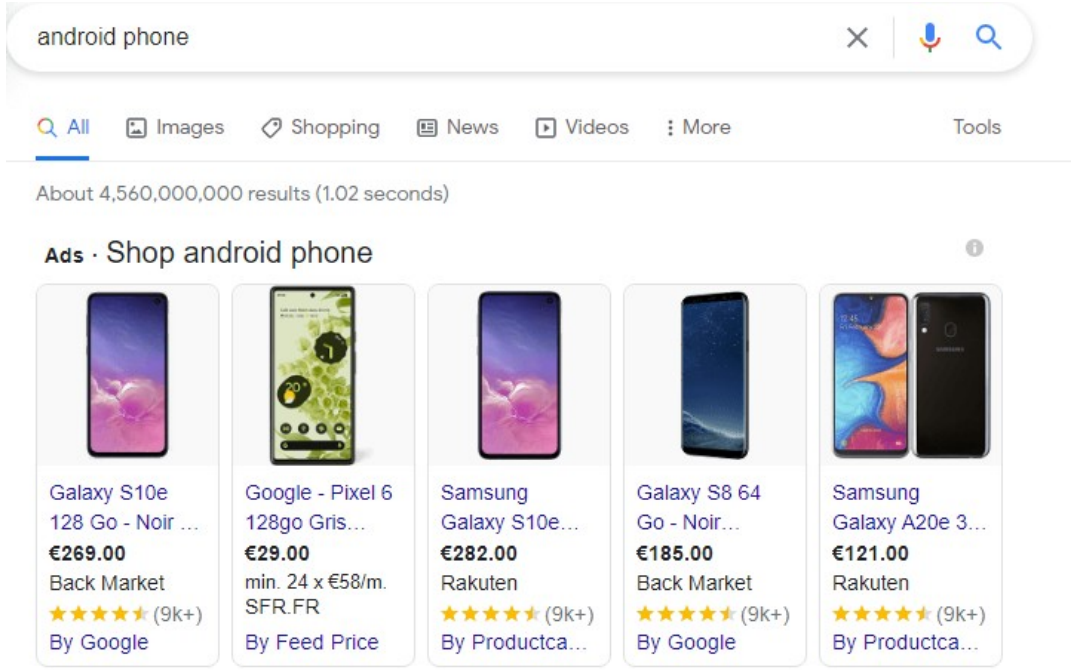


Figure 1: On the left, what browsers see. On the right, what humans see. Can we bridge the gap so that browsers see more of what we see?

RDFa Resource Description Framework in Attributes

See RDFa 1.1 Primer - Rich Structured Data Markup for Web Documents - https://www.w3.org/TR/rdfa-primer/

The Result



Google can extract structured data from e.g., e-commerce websites, and "understand" its meaning

Data Interoperability and Semantics

< Part 3. Data schemas and semantics >

Part 3.2. Semantics

Part 3.2.5. JSON-LD: JSON for Linking Data

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data

JavaScript Object Notation for Linked Data

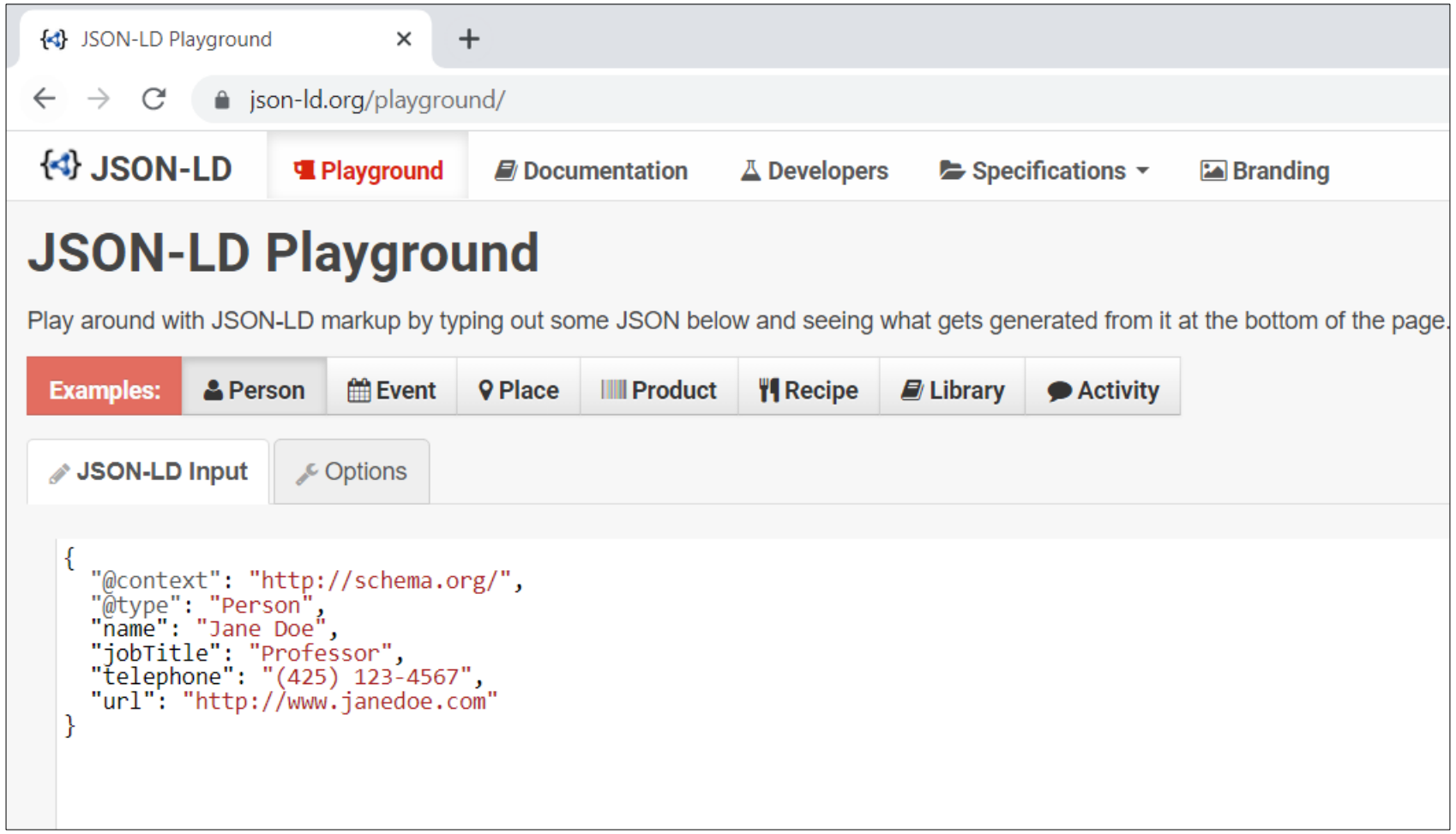
See JSON-LD 1.1 - A JSON-based Serialization for Linked Data - https://www.w3.org/TR/json-ld11/

JSON-LD

Filename extension	.jsonld
Internet media type	application/ld+json
Type of format	Semantic Web
Container for	Linked Data
Extended from	JSON
Standard	JSON-LD 1.1 / JSON-LD 1.1 API
Open format?	Yes

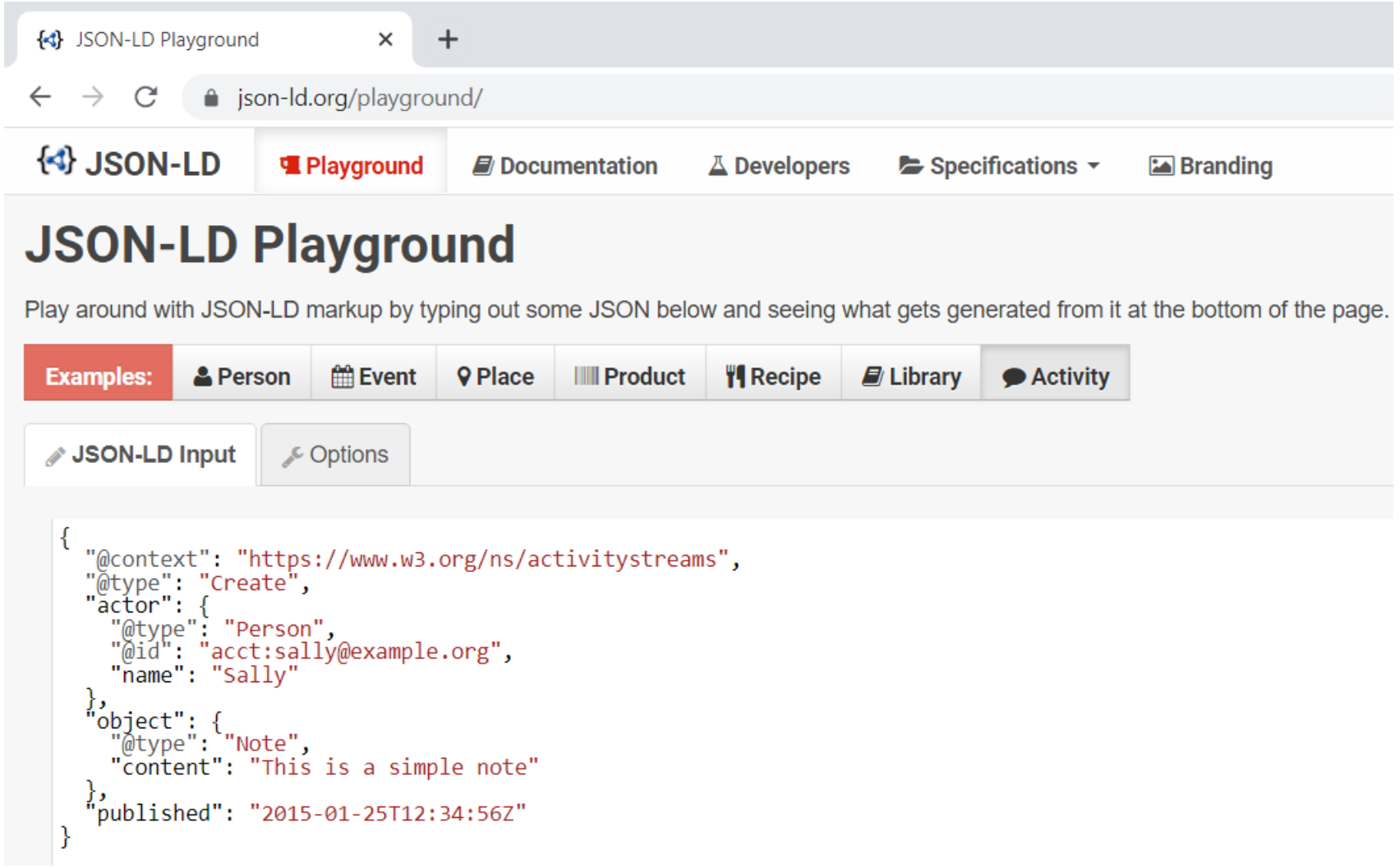
https://json-ld.org/learn.html

JavaScript Object Notation for Linked Data



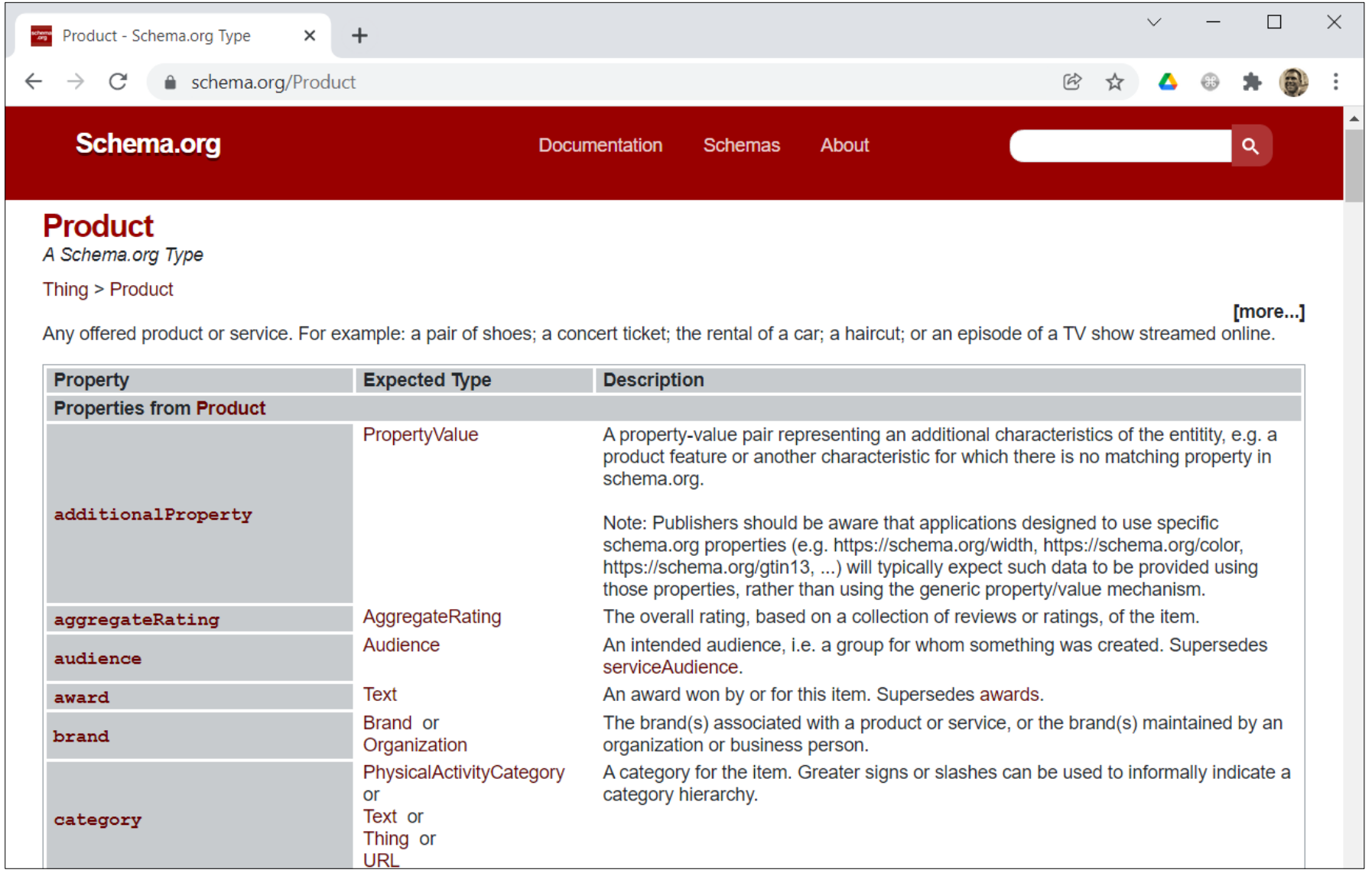
https://json-ld.org/playground/

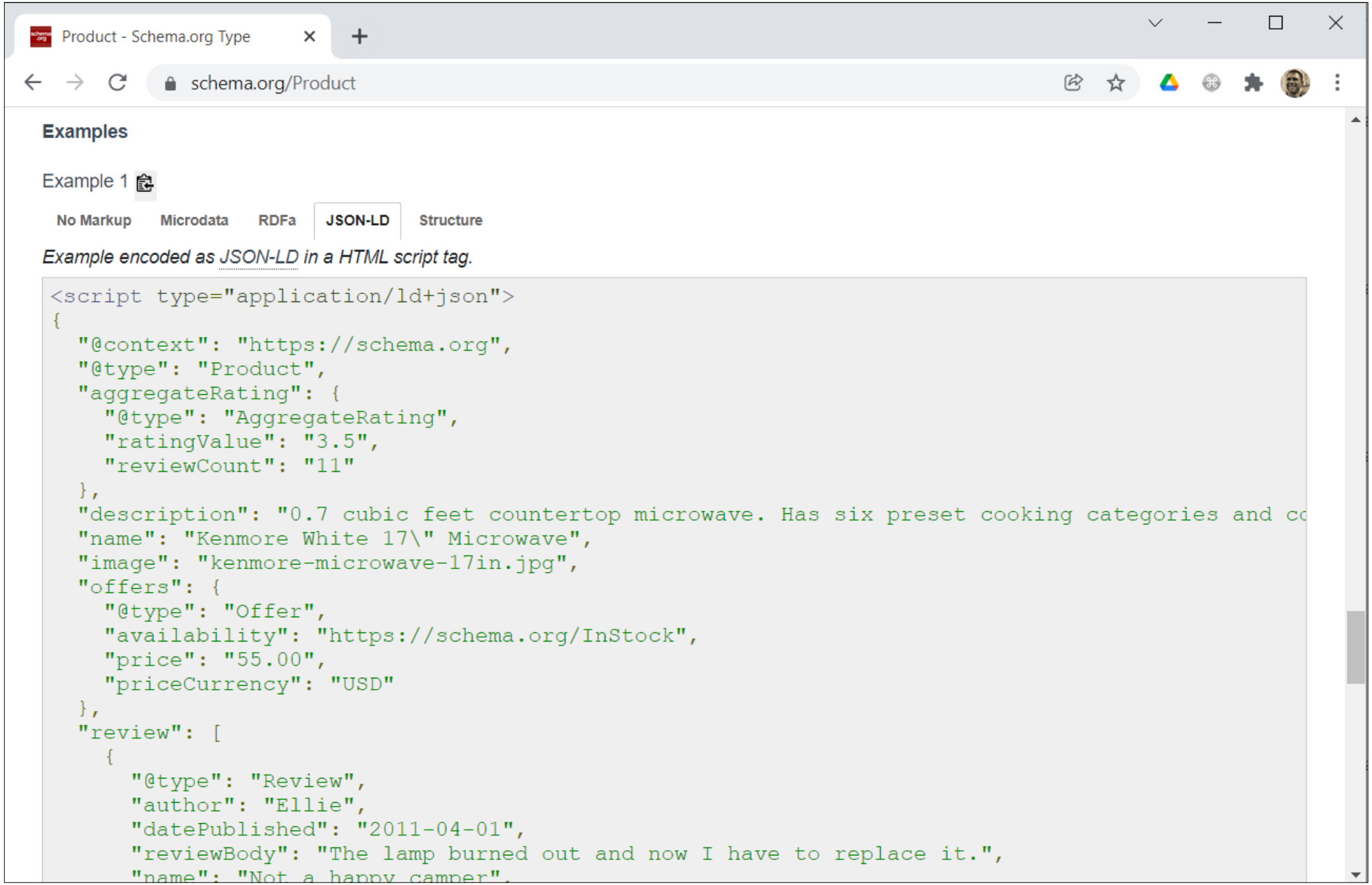
JavaScript Object Notation for Linked Data



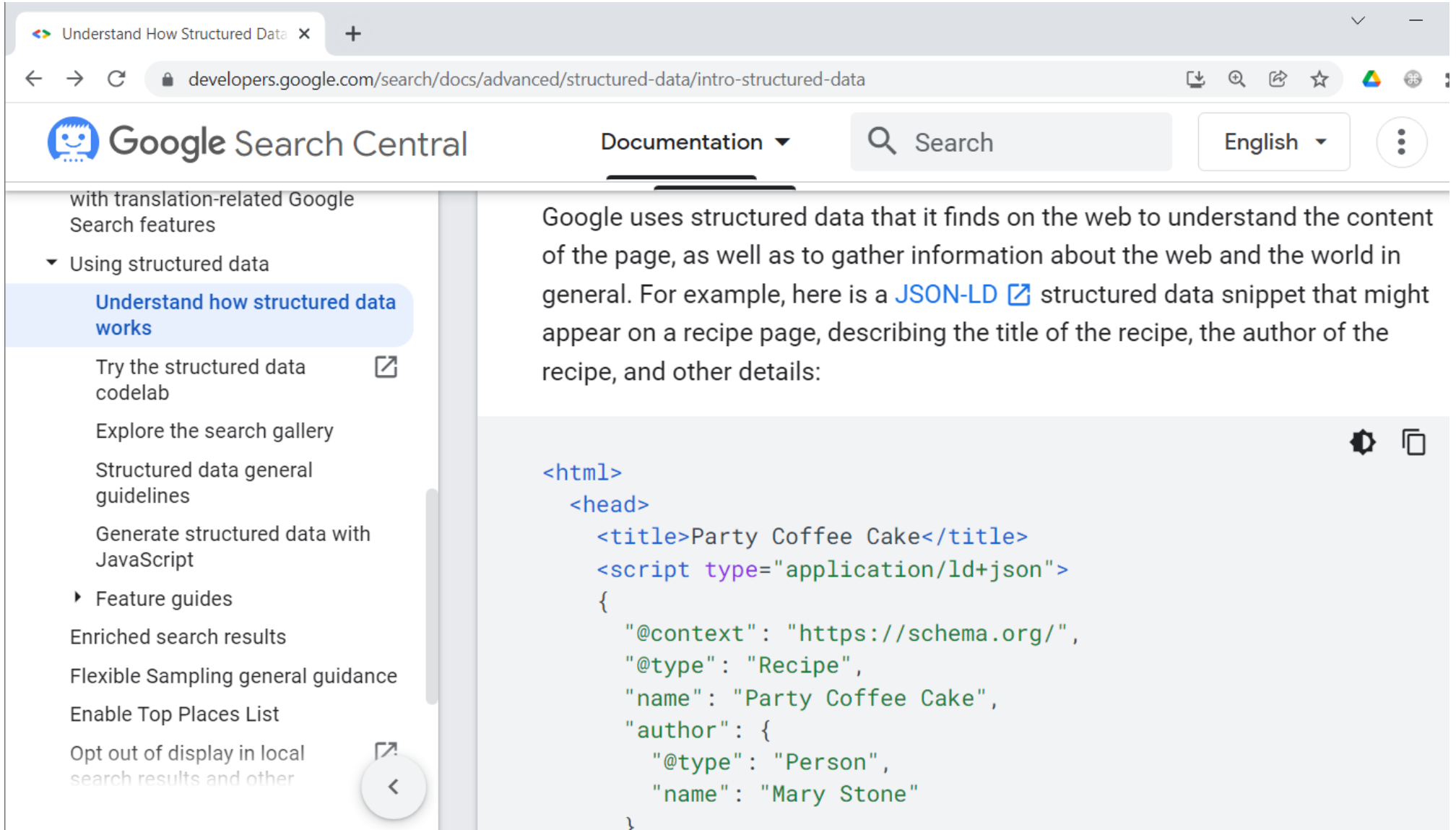
https://json-ld.org/playground/

JSON-LD and schema.org

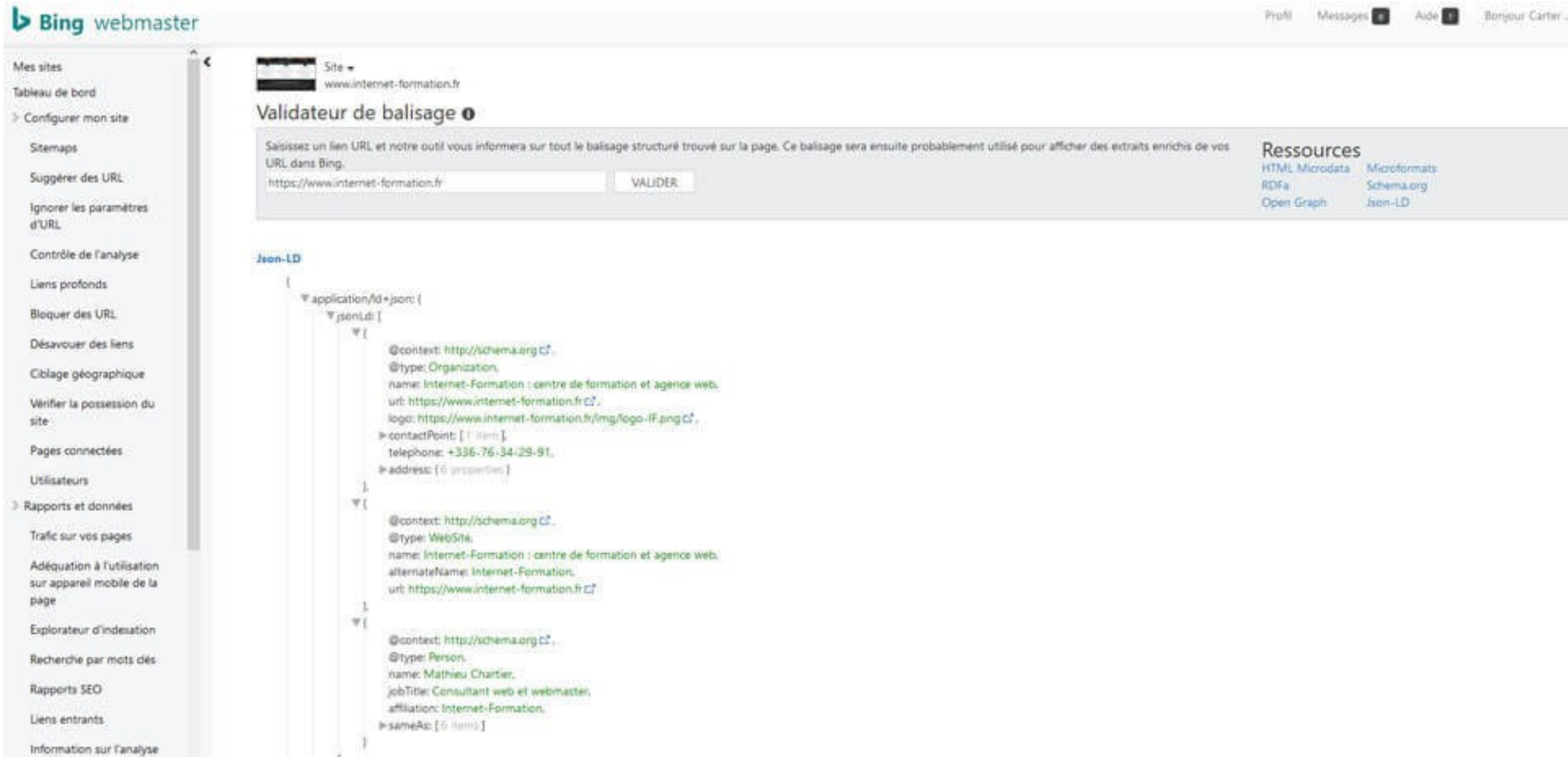



Product - Schema.org Type
schema.org/Product
Examples
Example 1
No Markup
Microdata
RDFa
JSON-LD
Structure
Example encoded as JSON-LD in a HTML script tag.
<script type="application/ld+json">
{
"@context": "https://schema.org",
"@type": "Product",
"aggregateRating": {
"@type": "AggregateRating",
"ratingValue": "3.5",
"reviewCount": "11"
},
"description": "0.7 cubic feet countertop microwave. Has six preset cooking categories and co
"name": "Kenmore White 17\" Microwave",
"image": "kenmore-microwave-17in.jpg",
"offers": {
"@type": "Offer",
"availability": "https://schema.org/InStock",
"price": "55.00",
"priceCurrency": "USD"
},
"review": [
{
"@type": "Review",
"author": "Ellie",
"datePublished": "2011-04-01",
"reviewBody": "The lamp burned out and now I have to replace it.",
"name": "Not a happy camper",

JSON-LD and Google



JSON-LD and Bing



Data Interoperability and Semantics

</ Part 3. Data schemas and semantics >

ICM – Toolbox Engineering and Interoperability of Software Systems – Course unit on Data Interoperability and Semantics
M1 Cyber Physical and Social Systems – Course unit on Data Interoperability and Semantics
Maxime Lefrançois https://maxime-lefrancois.info
Course unit URL: https://ci.mines-stetienne.fr/cps2/course/data