University of Central Lancashire
UCLan

CO3409
Distributed Enterprise Systems

FAIR principles
Metadata schemas

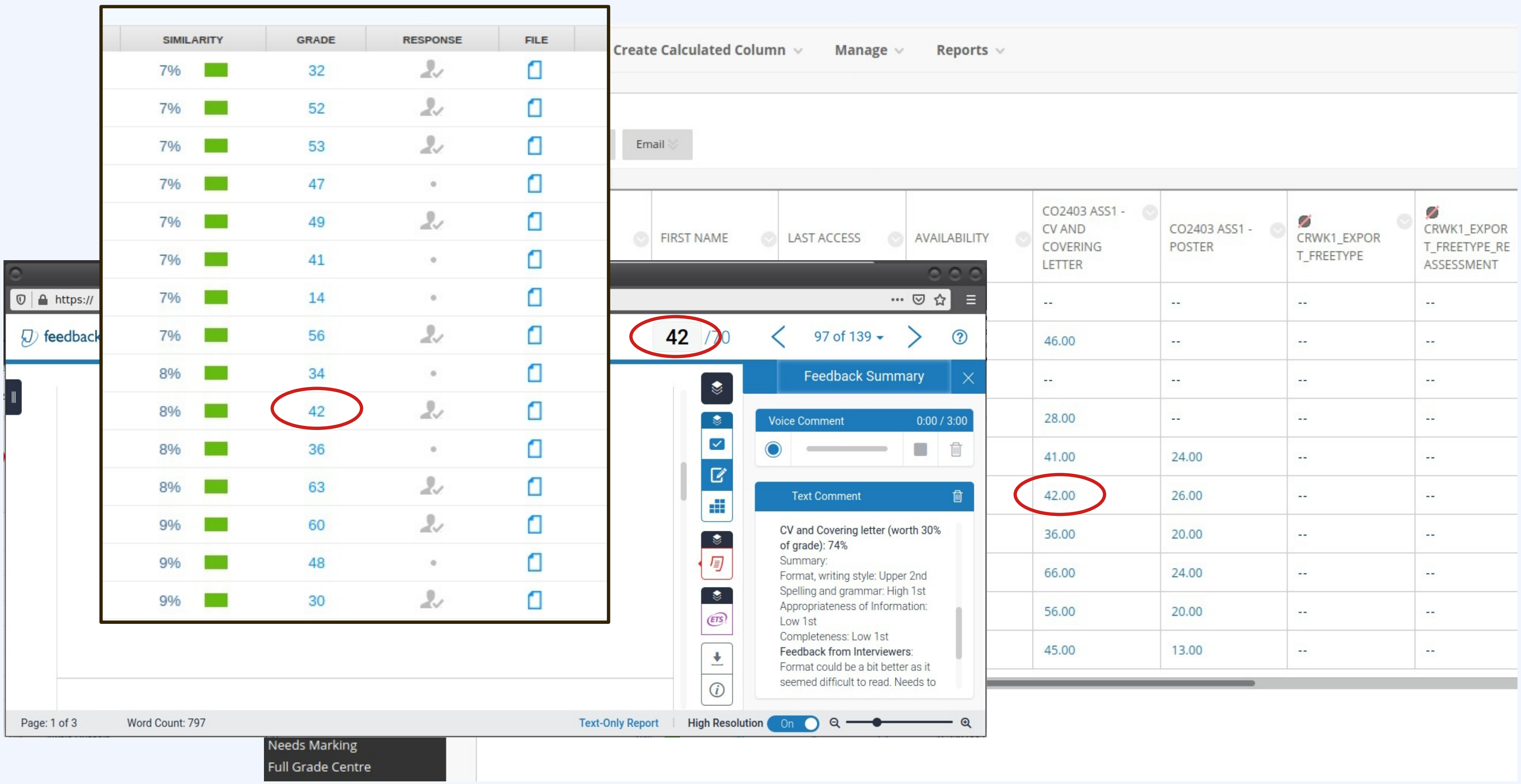
Where opportunity creates success

FAIR principles

CO3409

3rd March 2022

Data management & distributed enterprise systems



What do we (and the system) need to know to **use and reuse the data** correctly?

FAIR data and metadata

Metadata are "descriptive data about an object" (ISO 11179), usually a digital object

Digitization

data

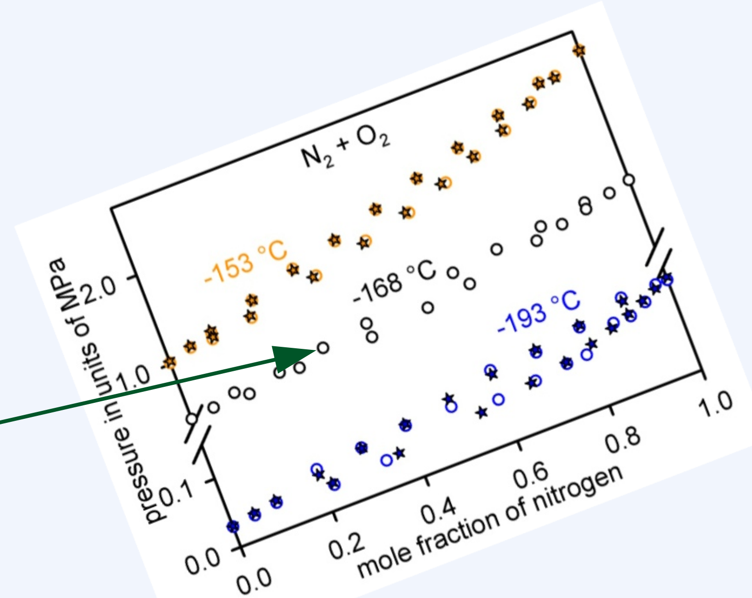
is about

Digitalization

Industry 5.0[2]

metadata
(data about data)

???



"How was the data point obtained?"
"How precise is it?" "Who did it?" …

Competency questions:[1]

Representative queries about data (*e.g.*, for metadata), to be competently answered by a knowledge base.

Good practice in data management (**FAIR princples**):

Make all data **findable**, **accessible**, **interoperable**, and **reusable**.

[1]M. Grüninger, M. S. Fox, in *Benchmarking: Theory and Practice*, doi:10.1007/978-0-387-34847-6_3, **1995**.
[2]M. Breque, L. De Nul, A. Petridis, *Industry 5.0*, EC policy brief, doi:10.2777/308407, **2021**.

University of Central Lancashire UCLan

FAIR principles of data management[1]

Findability

F1. Globally unique **persistent identifiers (PID)**
F2. **Enriched with metadata**
F3. Data identifier included in metadata
F4. **Registered**/indexed in **searchable platform**

Accessibility

A1. **Retrievable from PID** via a standard protocol
A1.1. Open and freely implementable protocol
A1.2. … **authentication/authorization** if necessary
A2. Metadata remain accessible (beyond data)

Interoperability

I1. **Formal language** used for **knowledge representation**
I2. Metadata use **vocabularies** that are themselves FAIR
I3. Semantic web principles, **data can refer to other data**

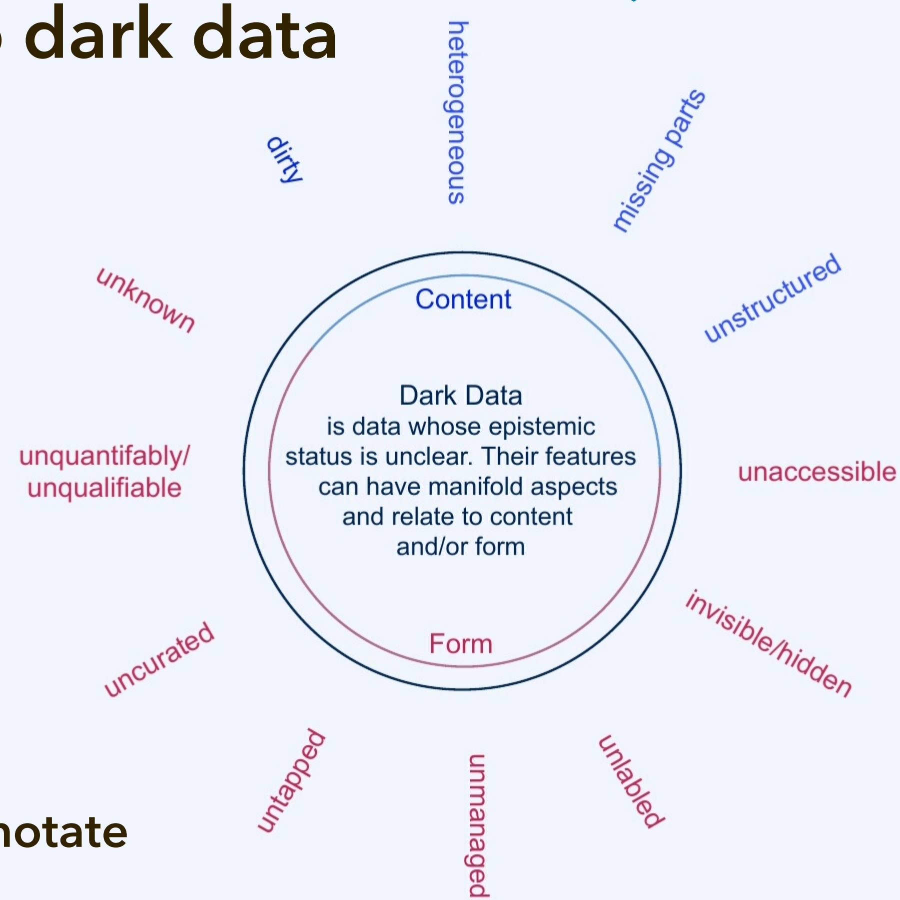
Reusability

R1. Metadata include a plurality of accurate and relevant attributes
R1.1. Release data and metadata with an accessible **data usage licence**
R1.2. Data are annotated with a detailed **provenance description**
R1.3. Relevant **disciplinary and community standards** are fulfilled

[1]M. D. Wilkinson *et al.*, "The FAIR Guiding Principles …," doi:10.1038/sdata.2016.18, **2016**.

FAIR data as opposed to dark data

To be FAIR, and therefore also **reusable**, the **epistemic status** (= knowledge status) of data needs to be characterized: Beyond interoperating via some I/O mechanisms, we must know in what way the data constitute knowledge.



(Figure: Courtesy B. Schembera.)

For data reusability, it is crucial to **annotate** data with all the required metadata.

The opposite of reusable data are **dark data**: Data with an uncharacterized epistemic status. Today there is a "deluge of dark data" – most data are dark.[1]

[1]B. Schembera, *J. Supercomput.* 77: 8946 – 8966, doi:10.1007/s11227-020-03602-6, **2021**.

Semantic interoperability

Three branches of the theory of formal languages:

- **Syntax**, theory of the **structure** of language → data formats (*e.g.*, JSON)
- **Semantics**, theory of the **meaning** of language → knowledge graph
- **Pragmatics**, theory of the **use** of language → processes and protocols

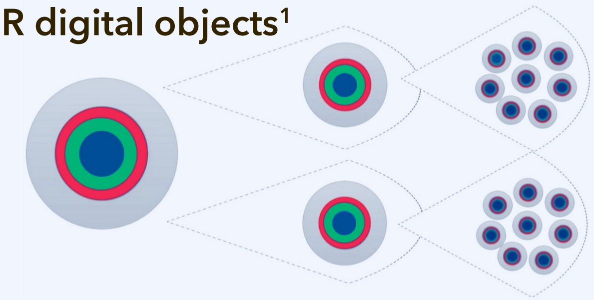
Generally speaking, semantics refers to "meaning," as opposed to syntax, which refers to "proper grammar and notation." Normally, there can only be a semantic content if there is a correct syntax, but the same content (*e.g.*, knowledge graph) can be represented in arbitarily many different formats.

Semantic interoperability allows us to coherently use many different formats, web service and API specifications, DB schemas and architectures. This is necessary whenever a distributed enterprise system is **internally heterogeneous**, or when it needs to **exchange information** with external systems.

University of Central Lancashire
UCLan

Semantic interoperability

FAIR digital objects[1]



Problems
Lack of (or overabundance of)

- P1: explicit definitions
- P2: common semantics (general ontologies)
- P3: reference repository
- P4: common metadata scheme across communities
- P5: metadata models

Recommendations

- R1: definitions of concepts, metadata and data schemes
- R2: creating semantic artefacts with open licenses
- R3: associated documentation for semantic artifacts
- R4: repositories of semantic artefacts
- R5: minimum metadata model and cross walks discovery
- R6: extensible options for disciplinary metadata
- R7: apply a broad definition of data (datasets, workflows, lab protocols, software, methods, hardware design, etc.)
- R8: clear protocols and building blocks for catalogues

Needs

- N1: principle approaches/tools for ontology and metadata schemes
- N2: harmonisation across disciplines
- N3: harmonisation of data of the same type
- N4: federated access to existing research data repositories

EUROPEAN OPEN SCIENCE CLOUD

EOSC Interoperability Framework[1]

[1]EOSC Executive Board, *EOSC Interoperability Framework*, doi:10.2777/620649, **2021**.

Semantic interoperability

```
# Model 1
# A       sigma          sigma_err
 40.0    1.17745        0.167
 60.0    3.03579        0.3592
 80.0    3.62384        0.3797
100.0    4.30474        0.3719


# Model 2
# A       sigma          sigma_err
 40.0    1.25022       0.1238
 60.0    2.75247       0.2723
 80.0    4.05209       0.2691
100.0    4.05401       0.2726
```

Now: How does this look in the context of concrete data?

Let us imagine we receive a file "sigma.dat" (on the right).

Three modes of interoperability (*i.e.*, agreements) are needed:

1) **Syntactic**: formal relations between signs → The file format (ASCII text file, tab separated columns, *etc.*).

2) **Semantic**: meaning, relations between signs and what they refer to → Info about the content: *e.g.*, what each column and block means, the data provenance, etc.

3) **Pragmatic**:[1, 2] relation between signs and their use, enviroment, users, and practices → *E.g.*: We type "rm sigma.dat" in a terminal. Depending on our rights on the file, it will be removed or not.

[1]The **EOSC Interoperability Framework** calls this **technical**, **organizational** and **legal** interoperability.
[2]On pragmatic interoperability for enterprise systems, see doi:10.1007/978-3-030-81200-3_4, **2021**.

Metadata schemas

CO3409

3rd March 2022

What do you see?

Use only simple sentences consisting of:

- A **subject**
- A **predicate**
- An **object**

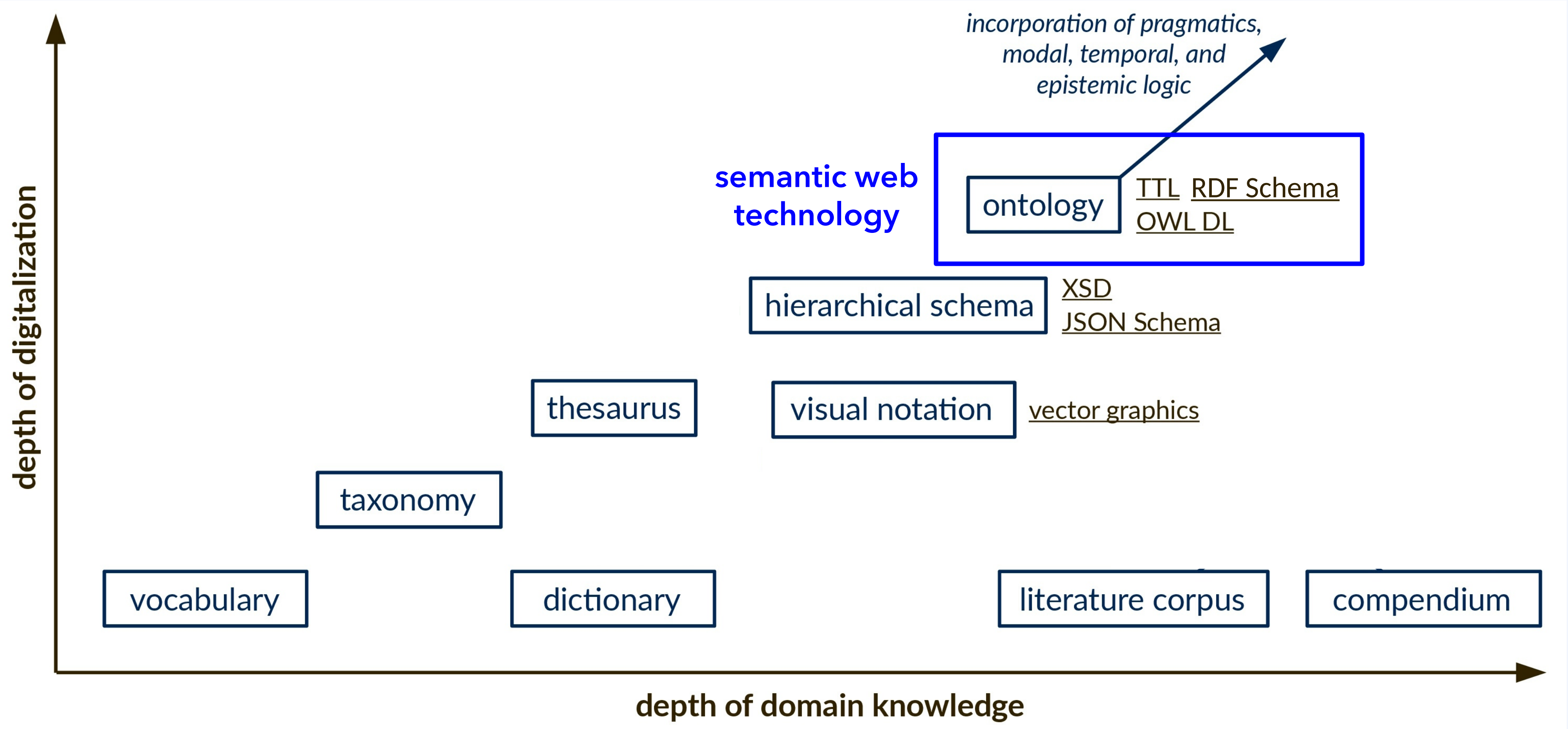
Such as:

"**The-elephant is-dancing-in the-room**."

"**The-wheel is-part-of the-car**."

Metadata standardization

Hierarchy of **semantic artefacts** (*i.e.*, metadata standards)

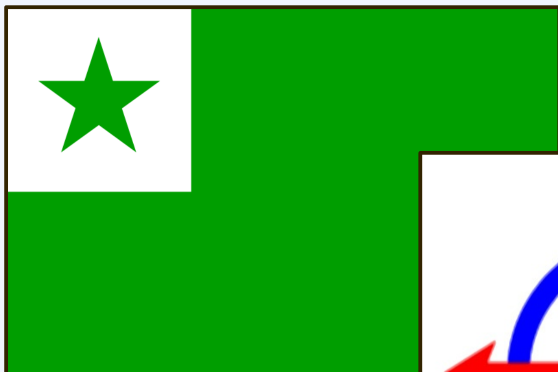


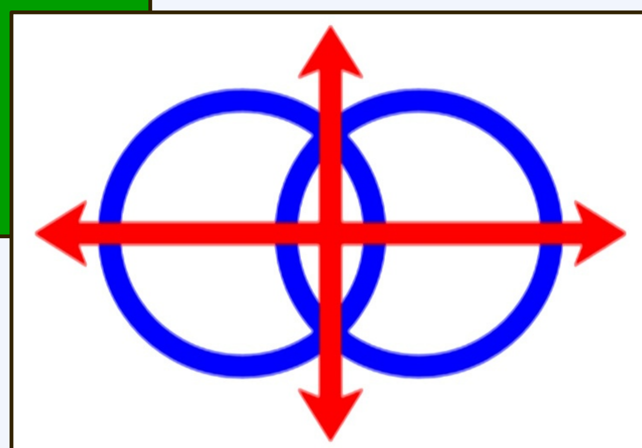
University of Central Lancashire
UCLan

Metadata standardization

„One World Language“

or now, “Web Ontology Lanugage (OWL),” based on RDF Schema (RDFS)





RDF Schema + Web Ontology Language

Ontologies are (data and) metadata **schemas for linked data**. They define what kinds of knowledge graphs are permitted. They specify what concepts can be instantiated by individuals, and what relations there can be between them; languages: **RDF Schema (RDFS)** and **Web Ontology Language (OWL)**.

animals:Fox a owl:Class.

concepts are of the type owl:Class

taxonomy specified using rdfs:subClassOf

animals:Fox rdfs:subClassOf animals:Canidae.
animals:Canidae rdfs:subClassOf animals:Mammalia.

animals:isNaturalEnemyOf a owl:ObjectProperty.

relations are of the type owl:ObjectProperty

animals:isNaturalEnemyOf rdfs:domain animals:Predator.

rdfs:domain and similarly **rdfs:range**

Schema/ontology design based on scenarios

What did you see?

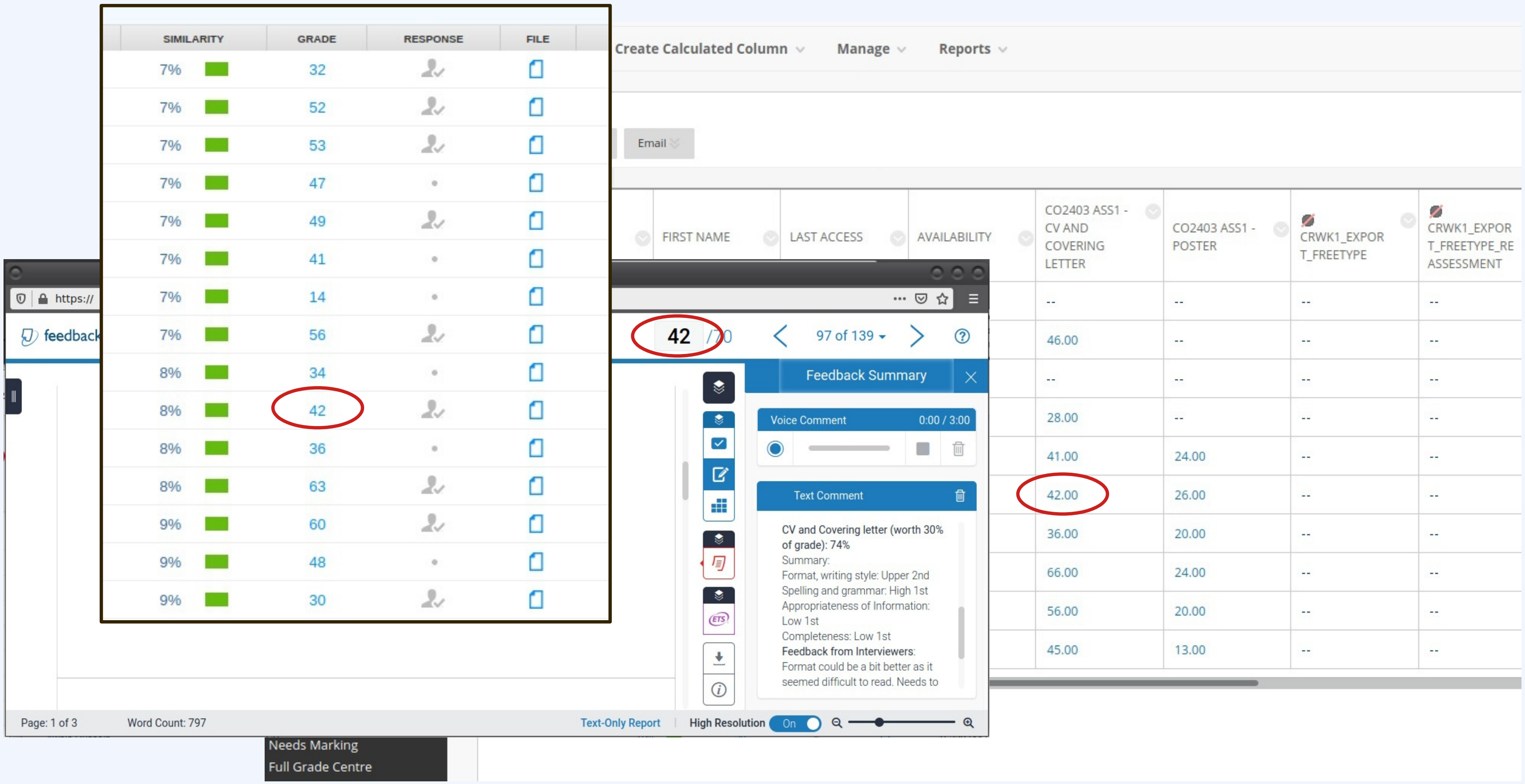
One approach to designing ontologies/schemas consists in **describing example scenarios**.

Usually, different people describe the same scenario in different terms, causing **semantic heterogeneity**.

Concepts instantiated by individuals should be in the ontology/RDF schema. **Relations** occurring as edges in the knowledge graph should be included; the **domains and ranges** of these relations should be included as concepts.

Different platforms may use different metadata schemas. To facilitate interoperability, an alignment is needed (*e.g.*, an **ontology alignment**).

Schema/ontology design via competency questions



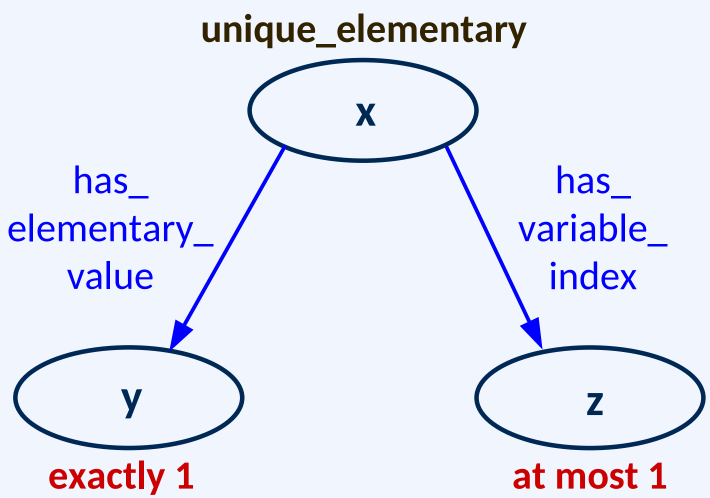
Another strategy for building an ontology consists in gathering **competency questions** and including the employed concepts and relations in the ontology.

Knowledge graph validation using SHACL

RDF schema, combined with the open world assumption, is very liberal in what knowledge graphs it permits. However, an API will usually need to specify a concrete kind of information content to be exchanged for a particular action.

Shapes Constraint Language (SHACL) can be used for such specifications.[1]

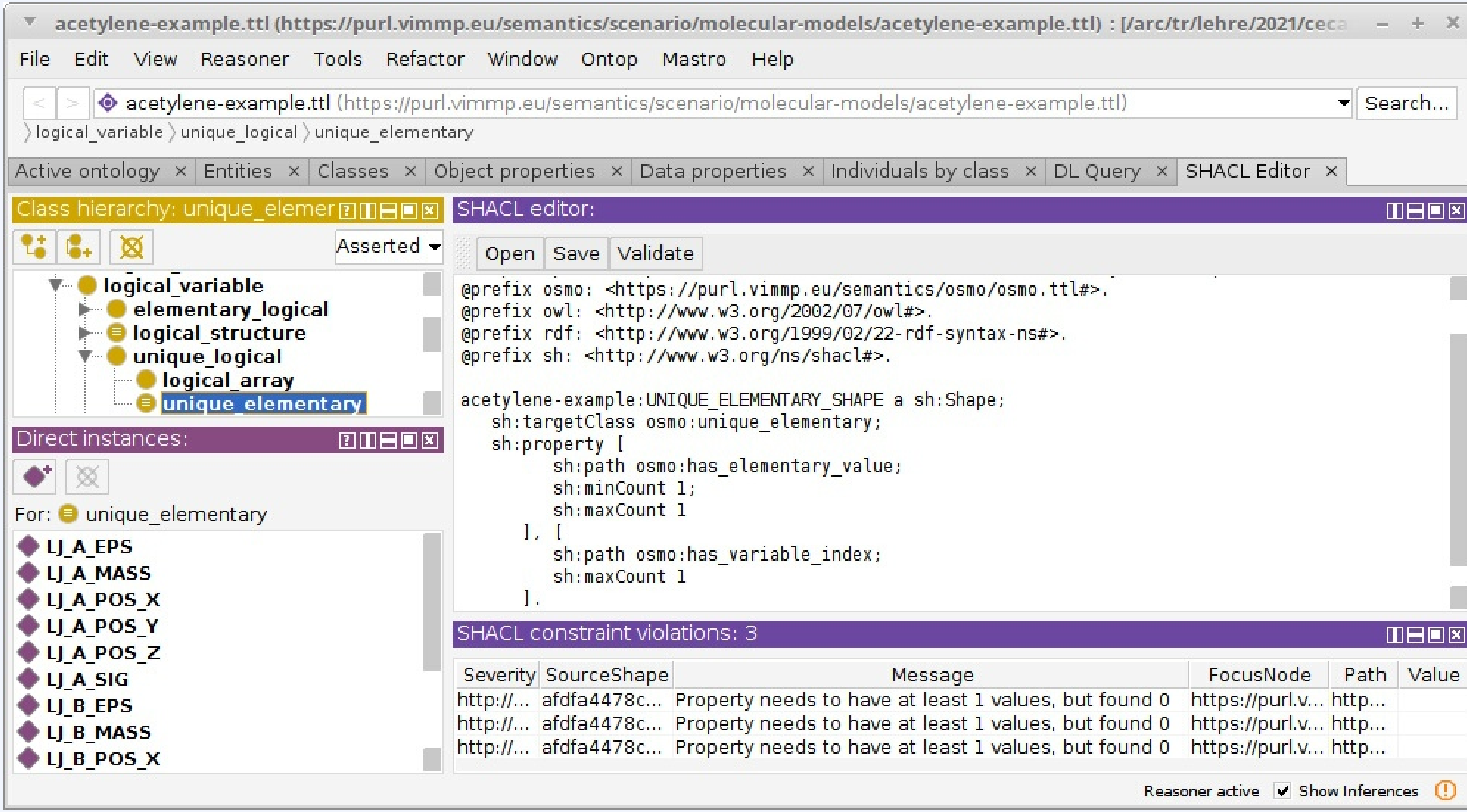
```
:unique_elementary_shape a sh:Shape;
  sh:targetClass :unique_elementary;
  sh:property [
      sh:path :has_elementary_value;
      sh:minCount 1;
      sh:maxCount 1
    ], [
      sh:path :has_variable_index;
      sh:maxCount 1
    ].
```



The open world assumption is **not** applied when evaluating SHACL constraints!

[1]W3C recommendation, https://www.w3.org/TR/shacl/, **2017**.

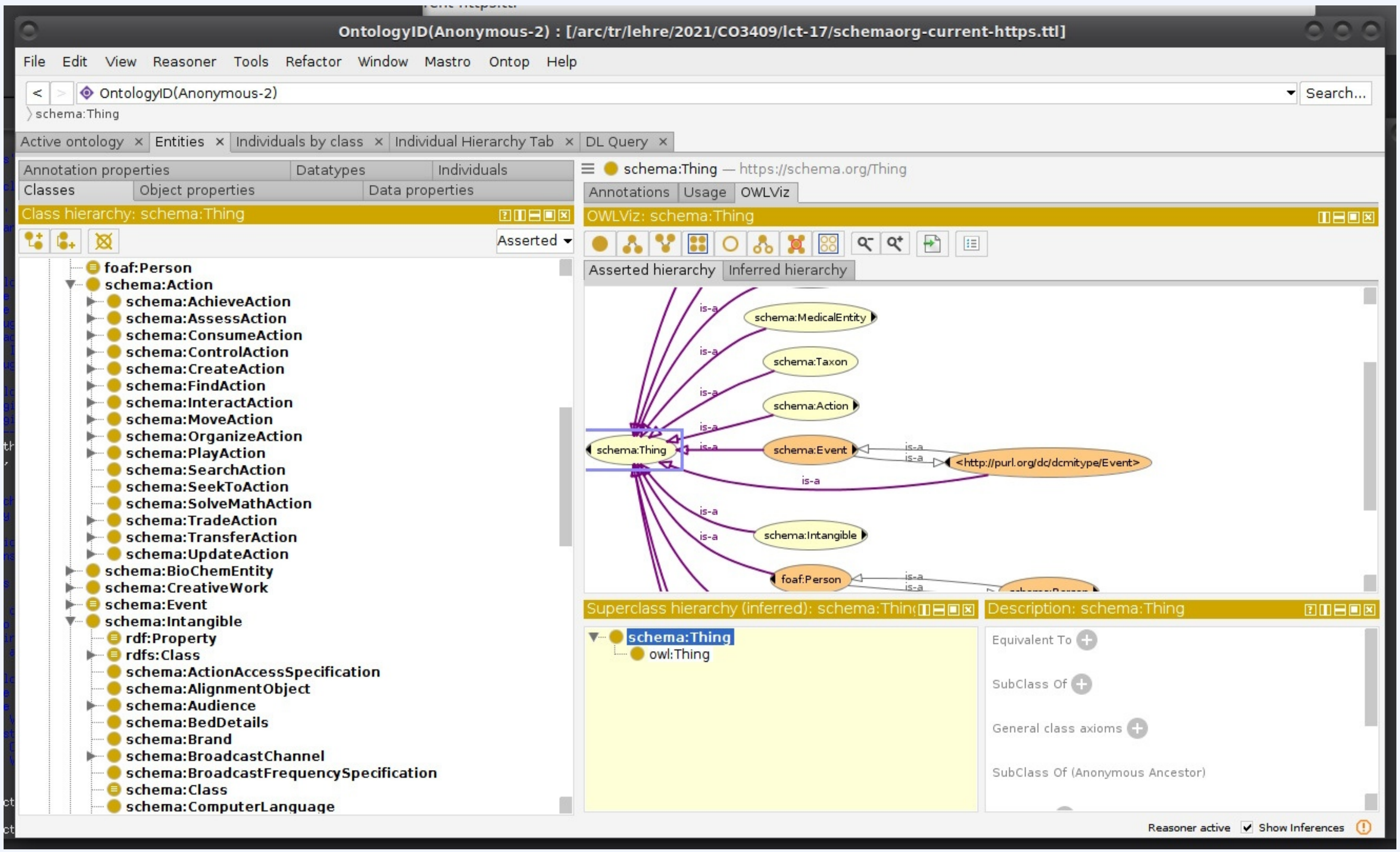
Knowledge graph validation using SHACL



(example: SHACL in Protégé)

Discussion

Schema.org: A metadata schema used by Google[1, 2]



[1]Schema.org definitions and documentation: https://schema.org/docs/full.html.

[2]Ontology in TTL format at https://schema.org/version/latest/schemaorg-current-https.ttl.

University of Central Lancashire
UCLan

CO3409 Distributed Enterprise Systems

FAIR principles
Metadata schemas

Where opportunity creates success