

# CO3409 Distributed Enterprise Systems

FAIR principles Metadata schemas

Where opportunity creates success



## FAIR principles

CO3409



### Data management & distributed enterprise systems

	SIMILARITY	GRADE	RESPONSE	FILE	Create Calculated Column v Manage v Reports v				
	7%	32	21	٥					
	7%	52	2.						
	7%	53	21	٥	Email 📎				
	7%	47		٥					
	7%	49	2,	٥	FIRST NAME LAST ACCESS AVAILABILITY	CO2403 ASS1 - CV AND	CO2403 ASS1 -	CRWK1_EXPOR	CRWK1_EXPOR
	7%	41		٥		COVERING	POSTER	T_FREETYPE	T_FREETYPE_RE ASSESSMENT
A https://	7%	14		0	🖂 🏠		12	12	
) feedback	7%	56	2.	٥	42 0 < 97 of 139 - > ⑦	46.00			
	8%	34		٥	Feedback Summary X			-	
	8%	42	2.	٥	S Voice Comment 0:00 / 3:00	28.00			12
	8%	36		٥		41.00	24.00		
	8%	63	2,	٥	Text Comment	42.00	26.00	11	11
	9%	60	2.	٥	CV and Covering letter (worth 30%	36.00	20.00		
	9%	48			of grade): 74% Summary: Format, writing style: Upper 2nd	66.00	24.00	12	2
	9%	30	2.		Spelling and grammar. High 1st Appropriateness of Information: Low 1st	56.00	20.00		
					Edw 1st     Completeness: Low 1st     Feedback from Interviewers:     Format could be a bit better as it	45.00	13.00		
					Format could be a bit better as it seemed difficult to read. Needs to				
age: 1 of 3	Word Count: 797			Text-	Dnly Report High Resolution On Q — Q				

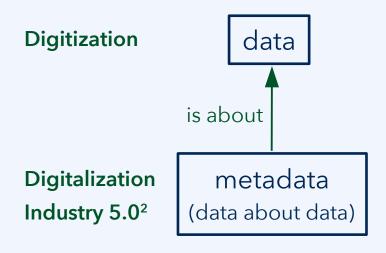
What do we (and the system) need to know to use and reuse the data correctly?

```
CO3409
```

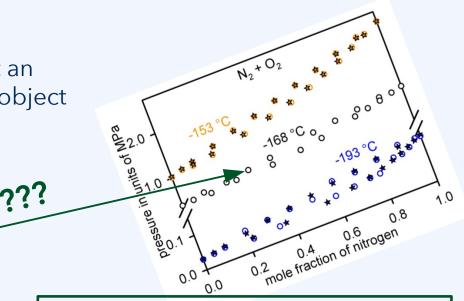


### FAIR data and metadata

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



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



### **Competency questions:**<sup>1</sup>

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.

<sup>1</sup>M. Grüninger, M. S. Fox, in *Benchmarking: Theory and Practice*, doi:10.1007/978-0-387-34847-6\_3, **1995**. <sup>2</sup>M. Breque, L. De Nul, A. Petridis, *Industry 5.0*, EC policy brief, doi:10.2777/308407, **2021**. CO3409 3<sup>rd</sup> March 2022 4



## FAIR principles of data management<sup>1</sup>

Findability

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

#### Interoperability

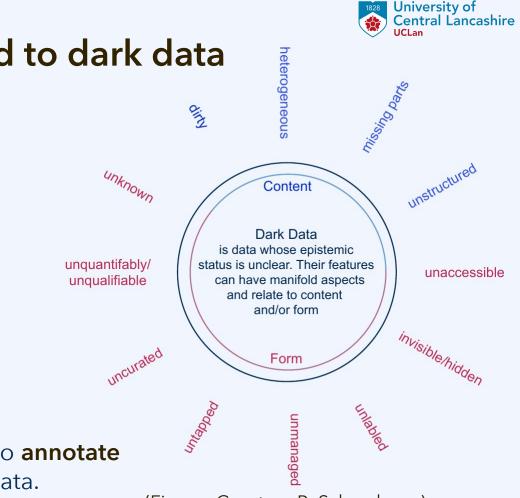
#### Accessibility

- 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)
- 11. Formal language used for knowledge representation
- 12. Metadata use vocabularies that are themselves FAIR
- 13. 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

<sup>1</sup>M. D. Wilkinson *et al.*, "The FAIR Guiding Principles ...," doi:10.1038/sdata.2016.18, **2016**. CO3409 3<sup>rd</sup> March 2022



### 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.

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

(Figure: Courtesy B. Schembera.)

6

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.<sup>1</sup> <sup>1</sup>B. Schembera, *J. Supercomput*. 77: 8946 – 8966, doi:10.1007/s11227-020-03602-6, **2021**. CO3409 3<sup>rd</sup> March 2022



### Semantic interoperability

Three branches of the theory of formal languages:

- **Syntax**, theory of the **structure** of language  $\rightarrow$  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 hetero-geneous, or when it needs to exchange information with external systems.



## Semantic interoperability





#### 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



#### Needs

N1: principle approaches/tools for ontology and metadata schemesN2: harmonisation across disciplinesN3: harmonisation of data of the same typeN4: federated access to existing research data repositories

#### 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



#### EOSC Interoperability Framework<sup>1</sup>

<sup>1</sup>EOSC Executive Board, EOSC Interoperability Framework, doi:10.2777/620649, **2021**.

CO3409



## Semantic interoperability

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

2) Semantic:

meaning, relations between signs and what they refer to

#### 3) Pragmatic:<sup>1,2</sup>

relation between signs and their use, enviroment, users, and practices

# 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

The file format

(ASCII text file, tab separated columns, etc.).

Info about the content: *e.g.*, what each column and block means, the data provenance, etc.

*E.g.*: We type "rm sigma.dat" in a terminal. Depending on our rights on the file, it will be removed or not.

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



### Metadata schemas

CO3409



### What do you see?



Use only simple sentences consisting of:

- A subject
- A predicate
- An object

Such as:

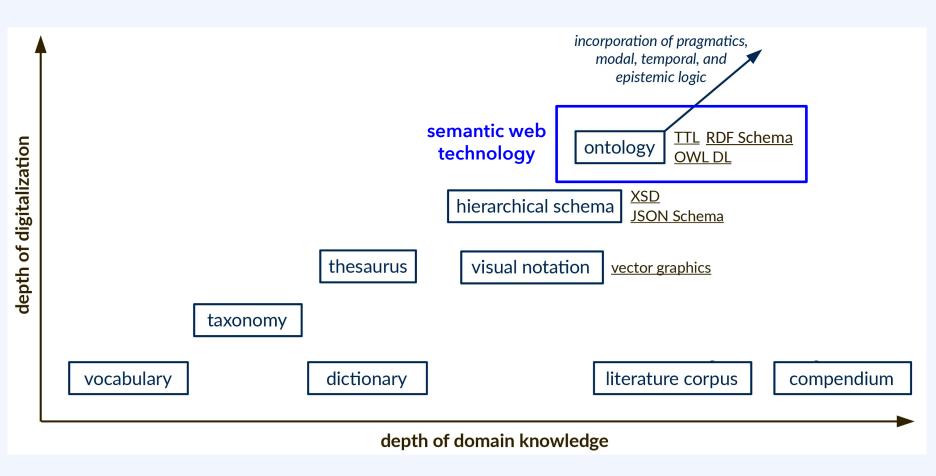
"The-elephant is-dancing-in the-room."

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



### Metadata standardization

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





### Metadata standardization

### "One World Language"





## 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)**.



### Schema/ontology design based on scenarios



### What did you see?

University of Central Lancashire

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**).

CO3409



### Schema/ontology design via competency questions

	SIMILARITY	GRADE	RESPONSE	FILE	Create Calculated Column 🗸 🛛 Manage 🗸 🛛 Reports 🗸	
	7%	32	2,	٥		
	7%	52	2,			
	7%	53	2,		Email 📎	
	7%	47				
	7%	49	2,	0		WK1_EXPC
	7%	41			COVERING POSTER T EREETVRE	FREETYPE_ SESSMENT
) 🔒 https://	7%	14		0	····· ··· ····························	
() feedback	7%	56	2.	0	<b>42</b> / 0 < 97 of 139 → ? ⑦ 46.00	
	8%	34		0	Feedback Summary X	
	8%	42	2,	٥	Voice Comment         0:00 / 3:00         28.00	
	8%	36		0		
	8%	63	2		Text Comment         42.00         26.00	
	9%	60	2.	٥	CV and Covering letter (worth 30% 36.00 20.00	
	9%	48		٥	Summary: Format, writing style: Upper 2nd 66.00 24.00	
	9%	30	2.	0	Spelling and grammar. High 1st Appropriateness of Information: Low 1st 56.00 20.00	
					Completeness: Low 1st Completeness: Low 1st Feedback from Interviewers: Format could be a bit better as it 45.00 13.00	
					Format could be a bit better as it     seemed difficult to read. Needs to	
age: 1 of 3	Word Count: 797			Text	Dnly Report High Resolution On Q — Q	

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

CO3409



## 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.<sup>1</sup>

```
:unique_elementary_shape a sh:Shape;
                                                       unique_elementary
 sh:targetClass :unique_elementary;
                                                              Χ
 sh:property [
                                                has
                                                                        has
     sh:path :has_elementary_value;
                                             elementary
                                                                      variable
                                                value
                                                                        index
     sh:minCount 1;
     sh:maxCount 1
                                                   γ
                                                                          Ζ
   ], [
                                               exactly 1
                                                                      at most 1
     sh:path :has_variable_index;
                                       The open world assumption is not applied
     sh:maxCount 1
                                          when evaluating SHACL constraints!
   ].
<sup>1</sup>W3C recommendation, https://www.w3.org/TR/shacl/, 2017.
```

3<sup>rd</sup> March 2022

CO3409



### Knowledge graph validation using SHACL

acetylene-example.ttl (https://purl.vimn	np.eu/semantics/scenario/molecular-models/acetylene-example.ttl) : [/arc/tr/lehre/2021/ceca –	+ ×
File Edit View Reasoner Tools Refact	or Window Ontop Mastro Help	
>logical_variable > unique_logical > unique_element		ch
Active ontology × Entities × Classes × C	Object properties × Data properties × Individuals by class × DL Query × SHACL Editor ×	
Class hierarchy: unique_elemer 🗈 🛙 🗖 🗖 🗷	SHACL editor:	
Asserted ▼	Open Save Validate	
Image: constraint of the second structure   Image: constraint of the second structure <th><pre>@prefix osmo: <https: osmo="" osmo.ttl#="" purl.vimmp.eu="" semantics="">. @prefix owl: <http: 07="" 2002="" owl#="" www.w3.org="">. @prefix rdf: <http: 02="" 1999="" 22-rdf-syntax-ns#="" www.w3.org="">. @prefix sh: <http: ns="" shacl#="" www.w3.org="">. acetylene-example:UNIQUE_ELEMENTARY_SHAPE a sh:Shape; sh:targetClass osmo:unique_elementary; sh:property [ sh:path osmo:has_elementary_value; sh:minCount 1; sh:maxCount 1 ], [ sh:path osmo:has_variable_index; sh:maxCount 1 ].</http:></http:></http:></https:></pre></th> <th>Ì</th>	<pre>@prefix osmo: <https: osmo="" osmo.ttl#="" purl.vimmp.eu="" semantics="">. @prefix owl: <http: 07="" 2002="" owl#="" www.w3.org="">. @prefix rdf: <http: 02="" 1999="" 22-rdf-syntax-ns#="" www.w3.org="">. @prefix sh: <http: ns="" shacl#="" www.w3.org="">. acetylene-example:UNIQUE_ELEMENTARY_SHAPE a sh:Shape; sh:targetClass osmo:unique_elementary; sh:property [ sh:path osmo:has_elementary_value; sh:minCount 1; sh:maxCount 1 ], [ sh:path osmo:has_variable_index; sh:maxCount 1 ].</http:></http:></http:></https:></pre>	Ì
LJ A POS Y		
♦ LJ_A_POS_Z		1
LJ_A_SIG	Severity         SourceShape         Message         FocusNode         Path         Y           http://         afdfa4478c         Property needs to have at least 1 values, but found 0         https://purl.v         https://purl.v	value
LJ_B_EPS	http:// afdfa4478c Property needs to have at least 1 values, but found 0 https://pull.v http	
◆ LJ_B_MASS ◆ LJ_B_POS_X	http:// afdfa4478c Property needs to have at least 1 values, but found 0 https://purl.v http	
	Reasoner active 🗸 Show Inference	s 🕕

(example: SHACL in Protégé)

### Discussion



University of Central Lancashire UCLan Concle 1, 2

Schema.org: A metadata schema used by Google<sup>1, 2</sup>

OntologyID(Anonymous-2) : [/	/arc/tr/lehre/2021/C03409/lct-17/schemaorg-current-https.ttl]
File Edit View Reasoner Tools Refactor Window Mastro Ontop Help	
< > OntologyID(Anonymous-2)	▼ Search
> schema: Thing	
Active ontology × Entities × Individuals by class × Individual Hierarchy Tab ×	DL Query ×
Annotation properties Datatypes Individuals	= eschema:Thing — https://schema.org/Thing
Classes Object properties Data properties	Annotations Usage OWLViz
Class hierarchy: schema:Thing	OWLViz: schema:Thing
🐮 🖡 🐹 Asserted 🗸	
6 foaf:Person	Asserted hierarchy
<ul> <li>schema:AssessAction</li> <li>schema:ConsumeAction</li> <li>schema:ConsumeAction</li> <li>schema:ControlAction</li> <li>schema:CreateAction</li> <li>schema:InteractAction</li> <li>schema:MoveAction</li> <li>schema:PlayAction</li> <li>schema:SearchAction</li> <li>schema:SearchAction</li> <li>schema:SearchAction</li> <li>schema:SearchAction</li> <li>schema:TradeAction</li> <li>schema:TradeAction</li> <li>schema:BioChemEntity</li> <li>schema:Event</li> </ul>	Superclass hierarchy (inferred): schema:Thing Description: schema:Thin
<ul> <li>schema:Intangible</li> <li>Grdf:Property</li> <li>Grdf:Class</li> <li>schema:ActionAccessSpecification</li> <li>schema:AulignmentObject</li> <li>schema:BedDetails</li> <li>schema:Brand</li> <li>schema:BroadcastChannel</li> <li>schema:BroadcastFrequencySpecification</li> <li>schema:BroadcastFrequencySpecification</li> </ul>	schema:Thing       Equivalent To +         owl:Thing       SubClass Of +         General class axioms +       SubClass Of (Anonymous Ancestor)
	Reasoner active 🗸 Show Inferences

<sup>1</sup>Schema.org definitions and documentation: https://schema.org/docs/full.html. <sup>2</sup>Ontology in TTL format at https://schema.org/version/latest/schemaorg-current-https.ttl. CO3409 3<sup>rd</sup> March 2022 20



# CO3409 Distributed Enterprise Systems

FAIR principles Metadata schemas

Where opportunity creates success