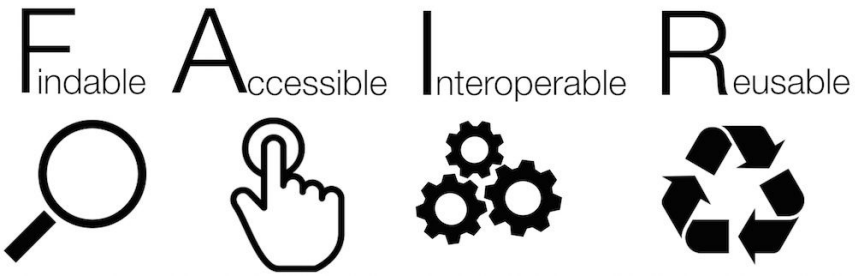
FAIR as a journey

Lessons learned and takeaways from building the GoTriple Discovery Platform for SSH

Luca De Santis - Net7

September 14th 2023 - PUBMET 2023, Zadar

FAIR as a goal



SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

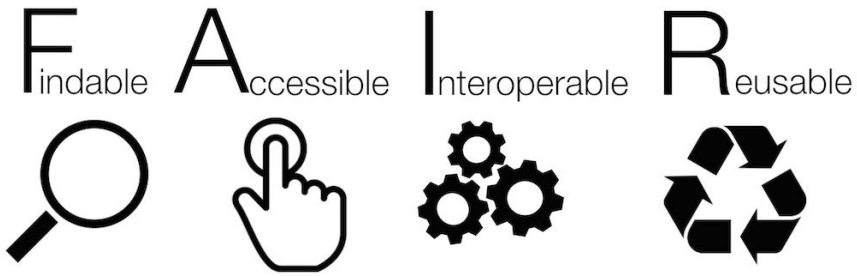
Mark D. Wilkinson *et al.*#

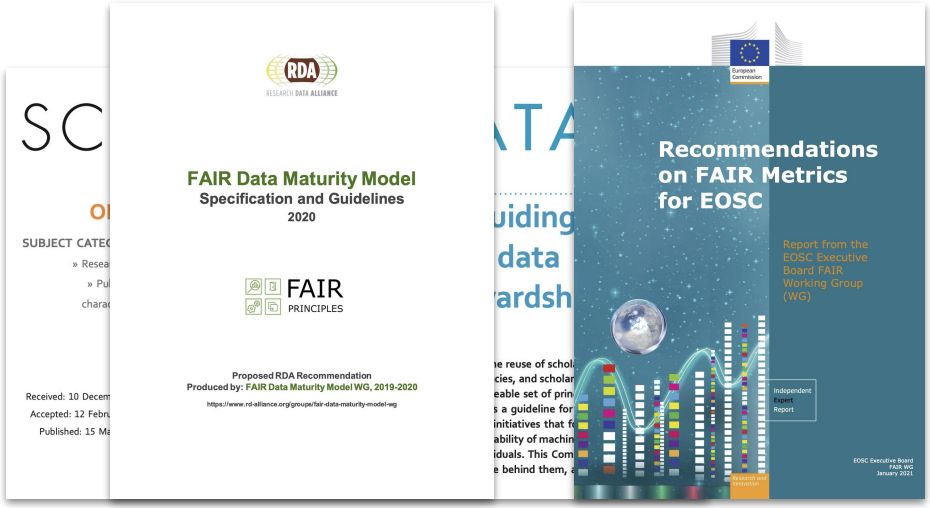
There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data

net7

FAIR as a goal





There is an urgent need to improve the infrastructure supporting the reuse of scholarly data

From “guiding principles to enable a full *exploitation* of data and their metadata”… … to something actionable

- FAIR Data Maturity Model: Indicators + Priorities + Evaluation Methods
- EOSC Metrics

net7

FAIR “as a journey”

- *FAIR should be seen as a journey*
- *Every community should define how to implement FAIR*
- *Inclusiveness: taking diversity into account*

net7

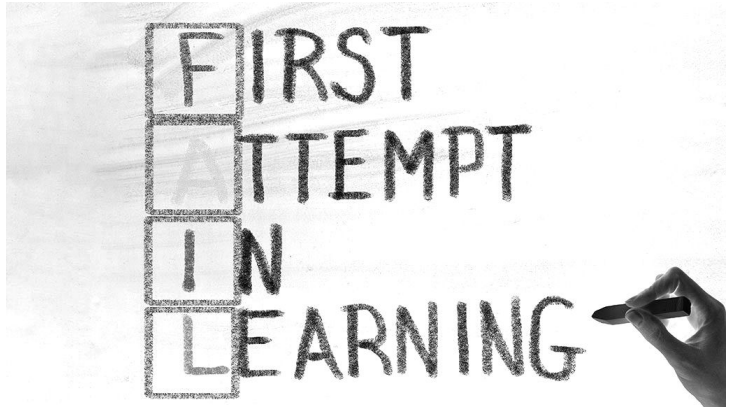
FAIR “as a journey”

Findable

Accessible

Interoperable

Reusable



- *Plan in advance, but…*
  - *…mistakes might happen! (and learn from them…)*

net7

Not only a *techie* thing

FAIR	ID	Indicator	Priority	
F1	RDA-F1-01M	Metadata is identified by a persistent identifier	●●●	Essential
F1	RDA-F1-01D	Data is identified by a persistent identifier	●●●	Essential
F1	RDA-F1-02M	Metadata is identified by a globally unique identifier	●●●	Essential
F1	RDA-F1-02D	Data is identified by a globally unique identifier	●●●	Essential
F2	RDA-F2-01M	Rich metadata is provided to allow discovery	●●●	Essential
F3	RDA-F3-01M	Metadata includes the identifier for the data	●●●	Essential
F4	RDA-F4-01M	Metadata is offered in such a way that it can be harvested and indexed	●●●	Essential
A1	RDA-A1-01M	Metadata contains information to enable the user to get access to the data	●●	Important
A1	RDA-A1-02M	Metadata can be accessed manually (i.e. with human intervention)	●●●	Essential
A1	RDA-A1-02D	Data can be accessed manually (i.e. with human intervention)	●●●	Essential
A1	RDA-A1-03M	Metadata identifier resolves to a metadata record	●●●	Essential
A1	RDA-A1-03D	Data identifier resolves to a digital object	●●●	Essential
A1	RDA-A1-04M	Metadata is accessed through standardised protocol	●●●	Essential
A1	RDA-A1-04D	Data is accessible through standardised protocol	●●●	Essential
A1	RDA-A1-05D	Data can be accessed automatically (i.e. by a computer program)	●●	Important
A1.1	RDA-A1.1-01M	Metadata is accessible through a free access protocol	●●●	Essential
A1.1	RDA-A1.1-01D	Data is accessible through a free access protocol	●●	Important
A1.2	RDA-A1.2-01D	Data is accessible through an access protocol that supports authentication and authorisation	●	Useful
A2	RDA-A2-01M	Metadata is guaranteed to remain available after data is no longer available	●●●	Essential
I1	RDA-I1-01M	Metadata uses knowledge representation expressed in standardised format	●●	Important
I1	RDA-I1-01D	Data uses knowledge representation expressed in standardised format	●●	Important
I1	RDA-I1-02M	Metadata uses machine-understandable knowledge representation	●●	Important
I1	RDA-I1-02D	Data uses machine-understandable knowledge representation	●●	Important
I2	RDA-I2-01M	Metadata uses FAIR-compliant vocabularies	●●	Important
I2	RDA-I2-01D	Data uses FAIR-compliant vocabularies	●	Useful
I3	RDA-I3-01M	Metadata includes references to other metadata	●●	Important
I3	RDA-I3-01D	Data includes references to other data	●	Useful
I3	RDA-I3-02M	Metadata includes references to other data	●	Useful

- Mostly technical requirements: can't be solved without IT.
- IT alone is not enough!
- Start by experimenting…

FAIR	ID	Indicator	Priority	
I3	RDA-I3-02D	Data includes qualified references to other data	●	Useful
I3	RDA-I3-03M	Metadata includes qualified references to other metadata	●●	Important
I3	RDA-I3-04M	Metadata include qualified references to other data	●	Useful
R1	RDA-R1-01M	Plurality of accurate and relevant attributes are provided to allow reuse	●●●	Essential
R1.1	RDA-R1.1-01M	Metadata includes information about the licence under which the data can be reused	●●●	Essential
R1.1	RDA-R1.1-02M	Metadata refers to a standard reuse licence	●●	Important
R1.1	RDA-R1.1-03M	Metadata refers to a machine-understandable reuse licence	●●	Important
R1.2	RDA-R1.2-01M	Metadata includes provenance information according to community-specific standards	●●	Important
R1.2	RDA-R1.2-02M	Metadata includes provenance information according to a cross-community language	●	Useful
R1.3	RDA-R1.3-01M	Metadata complies with a community standard	●●●	Essential
R1.3	RDA-R1.3-01D	Data complies with a community standard	●●●	Essential
R1.3	RDA-R1.3-02M	Metadata is expressed in compliance with a machine-understandable community standard	●●●	Essential
R1.3	RDA-R1.3-02D	Data is expressed in compliance with a machine-understandable community standard	●●	Important

net7

FAIR in GoTriple (step by step…)

About GoTriple.eu

OPERAS' **Multilingual discovery platform** for the SSH

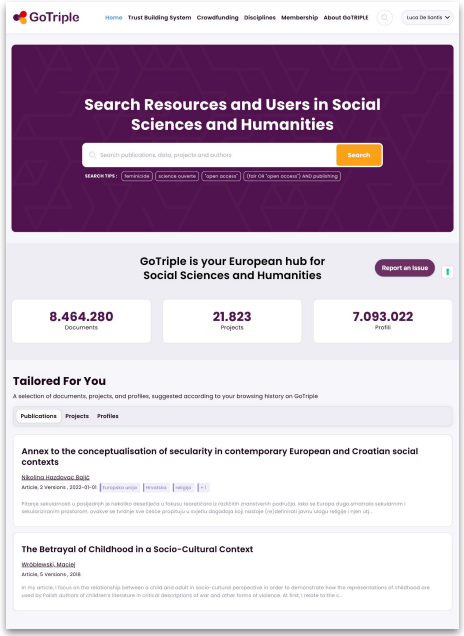
- The main outcome of the TRIPLE EU-funded project (Oct. 19 - March 23)

Search engine to discover Publications, Projects, Authors and Researchers Profiles

A **configurable harvesting and processing pipeline** to import and process publications and projects metadata

- ~12 million documents, 22.800 projects, 12 million authors… and counting
- Documents automatically harvested from small repositories and large aggregators alike
- Over 1.300 OAI-PMH endpoints managed (including Hrčak), dump imports from OpenAIRE, Isidore, CORDIS

Services for registered users! Please visit https://gotriple.eu



Findable & Accessible

- Findable
  - **Rich data model** for Documents, Authors, Projects
  - Harvested from multiple sources; their **identifiers maintained**
  - **Indexed** in a central search engine
- Accessible
  - **Standard access protocols**
  - Detailed explanations on the **origin of data**
  - **Interfaces** for **Humans & Machines**
  - **Metadata only** in the index. **Long term accessibility**.

To be Findable:

F1. (meta)data are assigned a globally unique and persistent identifier

F2. data are described with rich metadata (defined by R1 below)

F3. metadata clearly and explicitly include the identifier of the data it describes

F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1. the protocol is free, open and universally implementable

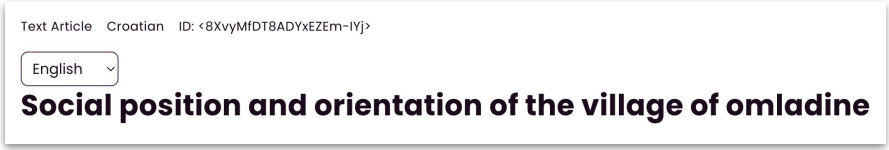
A1.2. the protocol allows for an authentication and authorization procedure, where necessary

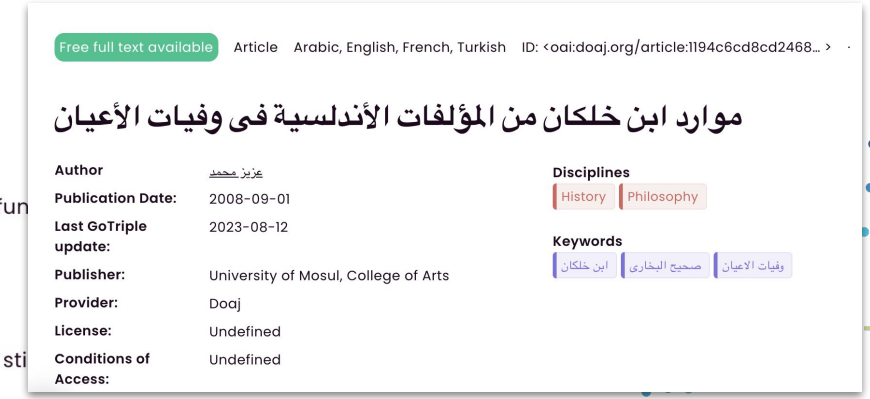
A2. metadata are accessible, even when the data are no longer available

(A Wider notion of) Accessibility…

***Not an "official" indicator** but quite important for a **wider accessibility** of GoTriple content*

- **We provide automatic translations in English when missing**!
  - Managed (for free!) thanks to the eTranslation EU service
- **Not bulletproof…** but good "enough"





Interoperable & Reusable

- Interoperable
  - Data model formalised via an **Ontology**
  - Expressive **FAIR-savvy vocabularies**
  - Use of **linked data references**
- Reusable
  - **License and conditions of access** preserved (*if present at the source*) and described in a **machine readable form**
  - **Data export** in multiple formats
  - REST **APIs**
  - **Documentation** available

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

I2. (meta)data uses vocabularies that follow FAIR principles

I3. (meta)data include qualified references to other (meta)data

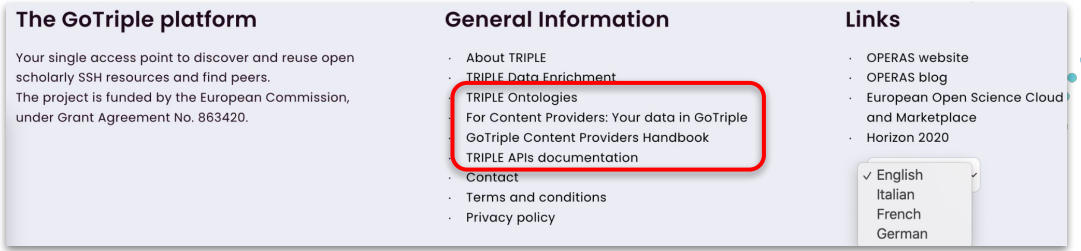
To be reusable:

R1. (meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (meta)data are released with a clear and accessible data usage license

R1.2. (meta)data are associated with data provenance

R1.3. (meta)data meet domain relevant community standards



The TRIPLE Ontologies

Born as an **experiment** at the end of TRIPLE!

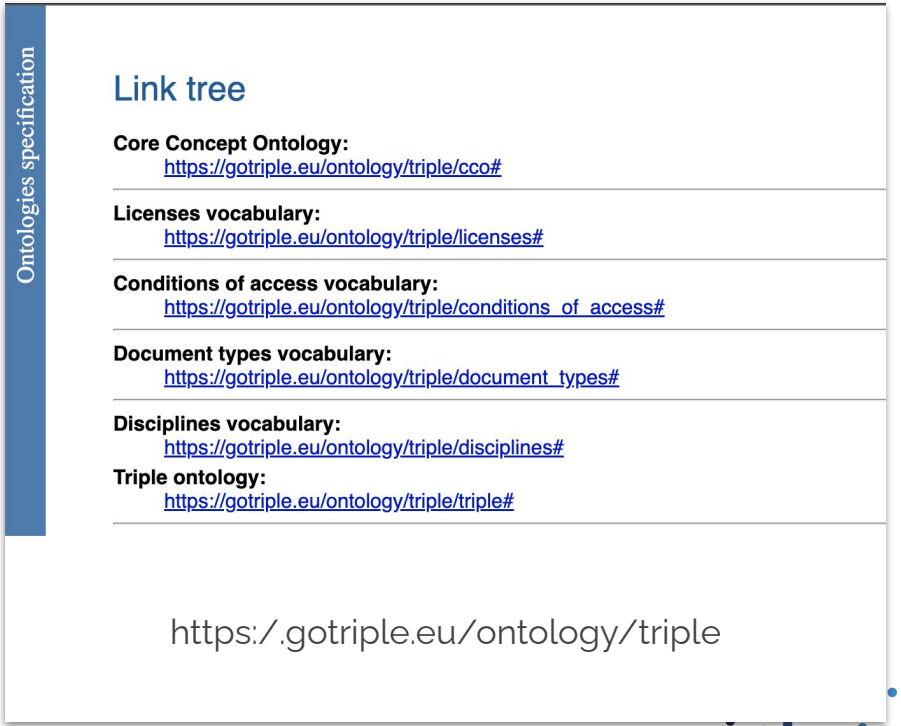
- A Master thesis of my colleague @Net7 **Alessandro Bertozzi**

Initial focus on **Documents**

- Born **to formalise and link the controlled vocabularies** used for documents enrichment
- Split into **multiple independent vocabularies** to **encourage reuse**

Created through an **iterative approach**.

- Currently in its 3rd iteration.
- Coming next: formalisation of **Profiles** and linking with Document entities; formalisation of **Projects**



Something to be reused beyond GoTriple

Classification of SSH disciplines

We proposed in TRIPLE **a classification for SSH** based on 27 disciplines.

The starting point was the taxonomy proposed in the **MORESS research project**

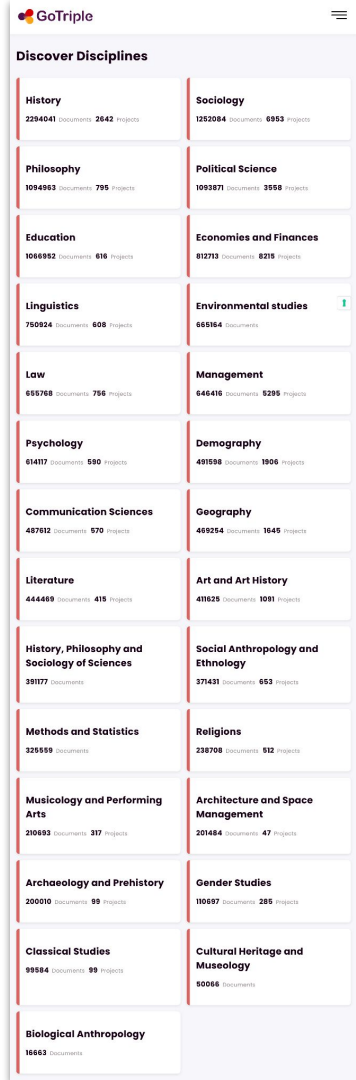
- *Mapping of research in European Social Sciences and Humanities*, 2003-2005

Reusable!

- It is used for classification in other OPERAS services

Not linked at the beginning!

https://gotriple.eu/disciplines



Classification of SSH disciplines

We proposed in TRIPLE **a classification for SSH** based on 27 disciplines.

The starting point was the taxonomy proposed in the **MORESS research project**

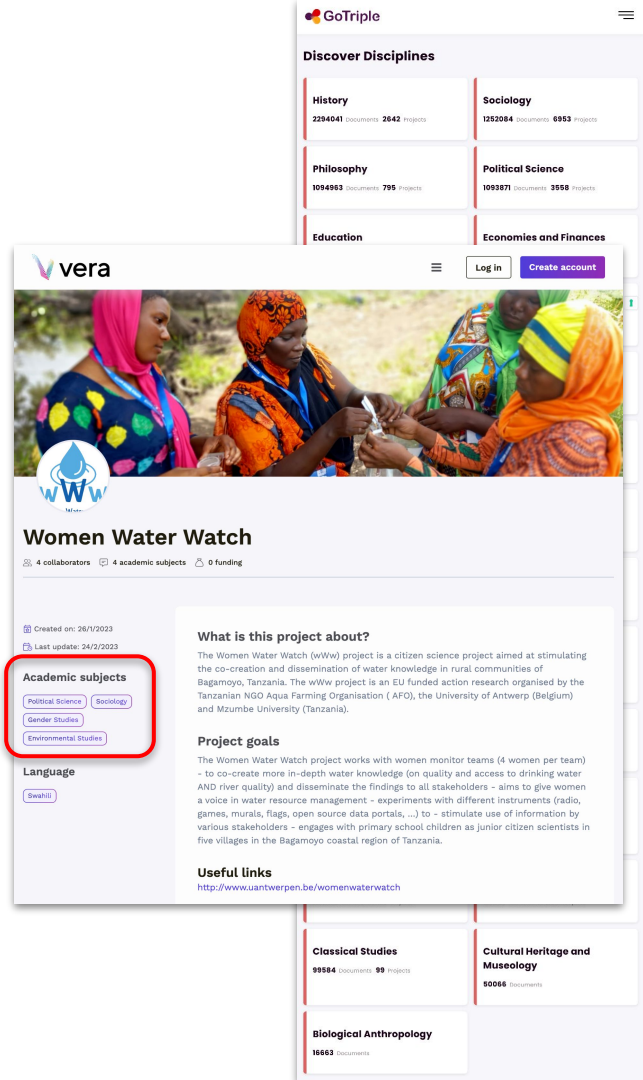
- *Mapping of research in European Social Sciences and Humanities*, 2003-2005

Reusable!

- It is used for classification in other OPERAS services (see vera.operas-eu.org)

Not linked at the beginning…

https://gotriple.eu/disciplines

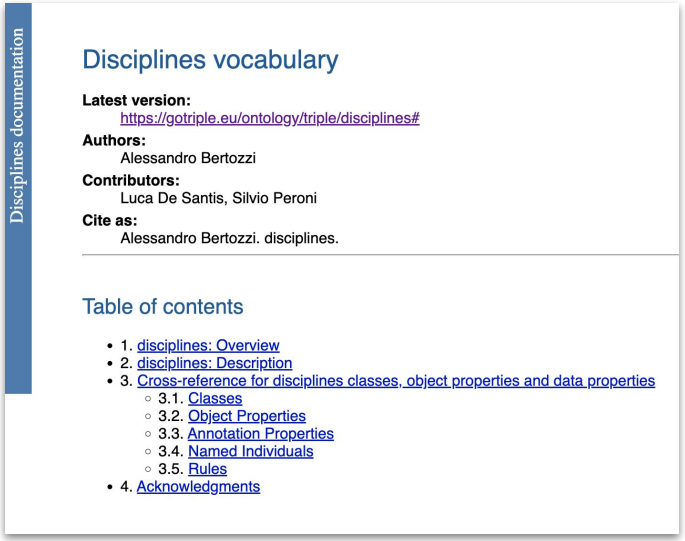


Classification of SSH disciplines

A formal SSH Disciplines vocabulary

- **SKOS** compliant
- **Multilingual** (English, Italian, French, German)
- **Linked** to multiple **standard ontologies** (Wikidata, LCSH, CESSDA, TRIPLE Vocabulary)

https://gotriple.eu/ontology/triple/disciplines


Disciplines documentation

Disciplines vocabulary

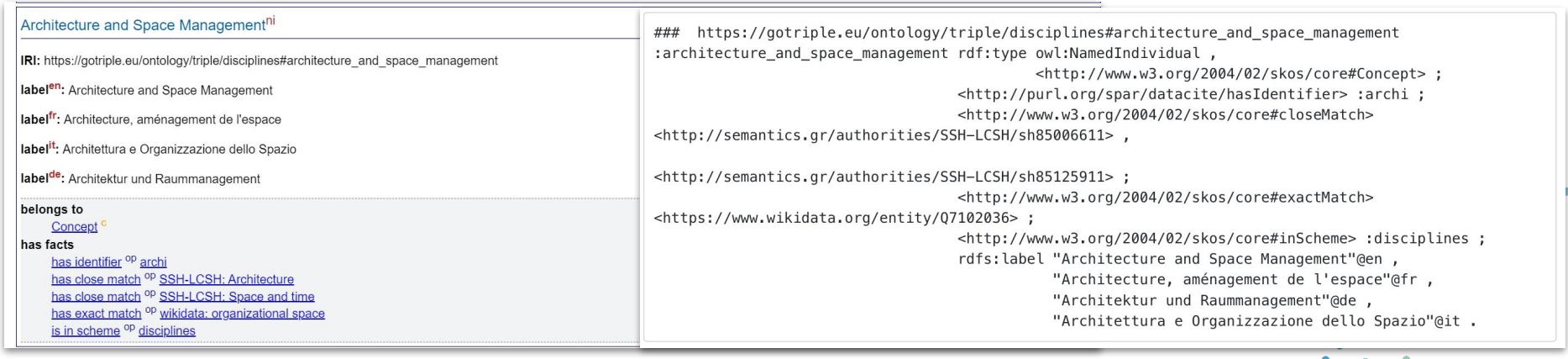
Latest version:
https://gotriple.eu/ontology/triple/disciplines#

Authors:
Alessandro Bertozzi

Contributors:
Luca De Santis, Silvio Peroni

Cite as:
Alessandro Bertozzi. disciplines.

Table of contents


Architecture and Space Management[ni]

IRI: https://gotriple.eu/ontology/triple/disciplines#architecture_and_space_management

label[en]: Architecture and Space Management

label[fr]: Architecture, aménagement de l'espace

label[it]: Architettura e Organizzazione dello Spazio

label[de]: Architektur und Raummanagement

belongs to
Concept [c]

has facts
has identifier [op] archi
has close match [op] SSH-LCSH: Architecture
has close match [op] SSH-LCSH: Space and time
has exact match [op] wikidata: organizational space
is in scheme [op] disciplines

```
###  https://gotriple.eu/ontology/triple/disciplines#architecture_and_space_management
:architecture_and_space_management rdf:type owl:NamedIndividual ,
                                            <http://www.w3.org/2004/02/skos/core#Concept> ;
                                   <http://purl.org/spar/datacite/hasIdentifier> :archi ;
                                   <http://www.w3.org/2004/02/skos/core#closeMatch>
<http://semantics.gr/authorities/SSH-LCSH/sh85006611> ,

<http://semantics.gr/authorities/SSH-LCSH/sh85125911> ;
                                   <http://www.w3.org/2004/02/skos/core#exactMatch>
<https://www.wikidata.org/entity/Q7102036> ;
                                   <http://www.w3.org/2004/02/skos/core#inScheme> :disciplines ;
                                   rdfs:label "Architecture and Space Management"@en ,
                                              "Architecture, aménagement de l'espace"@fr ,
                                              "Architektur und Raummanagement"@de ,
                                              "Architettura e Organizzazione dello Spazio"@it .
```

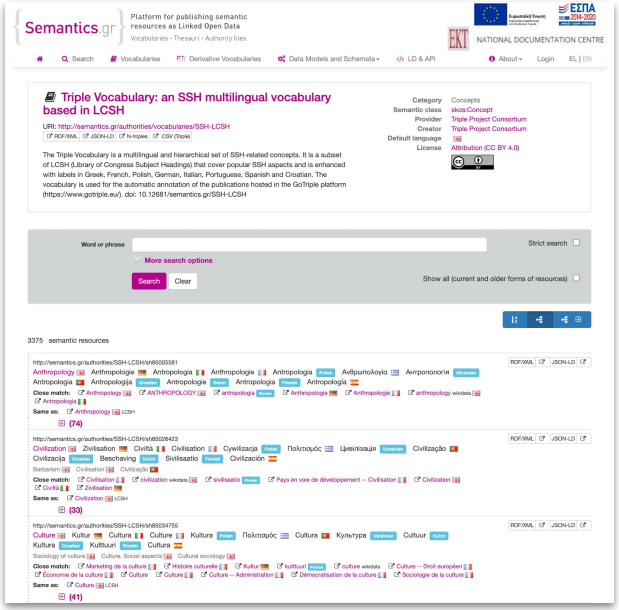
The TRIPLE Vocabulary

A multilingual **SKOS-based taxonomy** of concepts **related to SSH**

3.375 concepts, manually curated

Labels localised in **11+ languages**

Linked to the Library of Congress Subjects Headings (**LCSH**) vocabulary (*exactMatch*) and others, including Wikidata, CESSDA ELSST (*closeMatch*)



https://www.semantics.gr/authorities/vocabularies/SSH-LCSH/vocabulary-entries

Something that should have been done better...

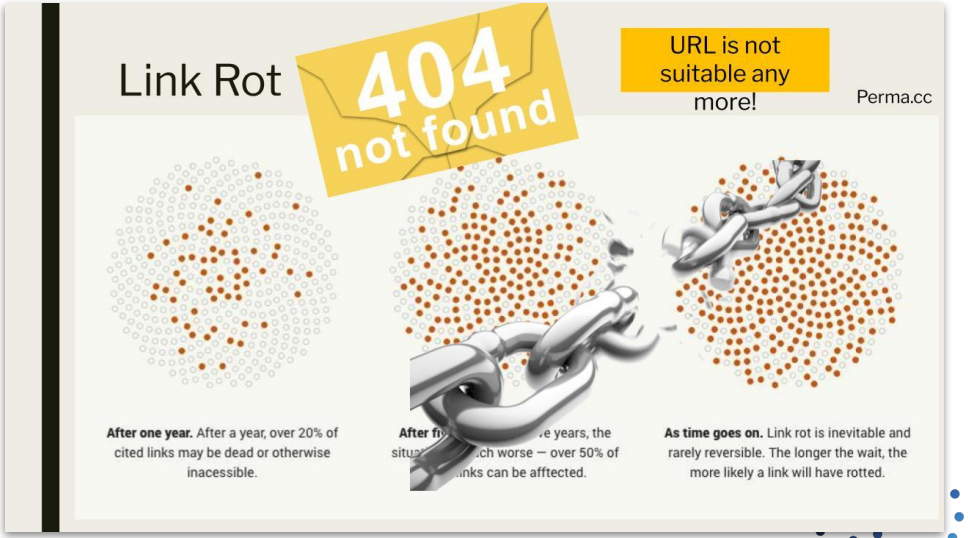
To PID or not to PID (for GoTriple…)

In GoTriple **the "official" Persistent Identifier** is…

…the **URI** of an entity (e.g. a document)!

In theory a good thing:

- **Unique**
- **Universal**
- **Persistent** (as long as GoTriple is alive…)
- Very **Semantic web**! :-)



In practice… **a bad idea…**

- "URL **is not suitable any more!**"
- **also a bad implementation**:
  https://www.gotriple.eu/document**s**/<ORIGINAL_ID>
- `https://gotriple.eu/documents/oai%3Adoaj.org%2Farticle%3A1194c6cd8cd24682b7b164c6df90c009`

To PID or not to PID (for GoTriple…)

Using the original main identifier as the basis for **generating** a **local ID**

…but **with a transformation**!

- See **OpenAIRE approach** (with *gotchas…*)

> OpenAIRE assigns internal identifiers for each object it collects. By default, the internal identifier is generated as `sourcePrefix::md5(localId)` where:
>
> - `sourcePrefix` is a namespace prefix of 12 chars assigned to the data source at registration time
> - `localid` is the identifier assigned to the object by the data source

Current GoTriple PIDs are **not so easily fixable**…

- **To be definitely addressed** in a continuation of GoTriple

On the other hand, generating "official PIDs" (via Handle.net, Datacite…) for GoTriple's entities…

- …not a good idea **for an aggregator** (see duplication of PIDs…)

A closing thought: a call to action to improve metadata quality

(Meta)Data (quality) is king!

OAI-PMH/DC is widely used for metadata publishing

Can we overcome its "**semantic simplicity**" *without disrupting the standard*?

- Find smart **common practices**

```
<dc:subject>Livestock</dc:subject>
<dc:subject xml:lang="sl-SI">slovenski jezik</dc:subject>
<dc:subject xsi:type="dcterms:LCC">Agriculture (General)</dc:subject>
```

```
<dc:creator ">Walk, Paul</dc:creator>
<dc:creator pid="https://orcid.org/0000-0003-1541-5631 http://paulwalk.net">Walk, Paul</dc:creator>
```

(Meta)Data (quality) is king!

OAI-PMH/DC is widely used for metadata publishing

Can we overcome its "**semantic simplicity**" *without disrupting the standard*?

- Find smart **common practices**
- Identify the right **Semantic Artefacts** for each scientific domain

Software might help!

- E.g. **plug-ins** for Journal and DAM systems: OJS, Lodel, Fedora Commons, Dspace,

net7

Thanks for your attention!

Luca De Santis

desantis@netseven.it
https://gotriple.eu/profile/ldesantis

Did you like it? Please share!

Image credits

- Slides 2, 3: “FAIR image” taken from https://www.panosc.eu/data/fair-principles/
- Slides 4, 5: “FAIR image” taken from https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-300.html
- Slide 5: “FAIL image” taken from https://www.snexplores.org/article/secret-science-mistakes-boost-understanding
- Slide 6: taken from RDA FAIR Data Maturity Model Specification and Guidelines 2020
- Slides 9, 11: taken from Recommendations on FAIR Metrics for EOSC
- Slide 17: taken from the TRIPLE Vocabulary home page https://www.semantics.gr/authorities/vocabularies/SSH-LCSH/vocabulary-entries
- Slide 19: taken from “PIDs in the SSH - Current state and upcoming challenges” by Jadranka Stojanovski, TRIPLE Booksprint, The role of open metadata in the SSH scholarly communication, Konstancin-Jeziorna, 7-9 September 2022
- Slide 20: taken from OpenAIRE Graph documentation “PIDs and identifiers” https://graph.openaire.eu/docs/5.1.2/data-model/pids-and-identifiers/
- Slides 22, 23: © Gary Waters/Getty Images/Ikon Images. Taken from https://www.npr.org/2017/03/23/521195903/how-the-scarcity-mindset-can-make-problems-worse

All other images taken from the OPERAS services GoTriple & VERA.

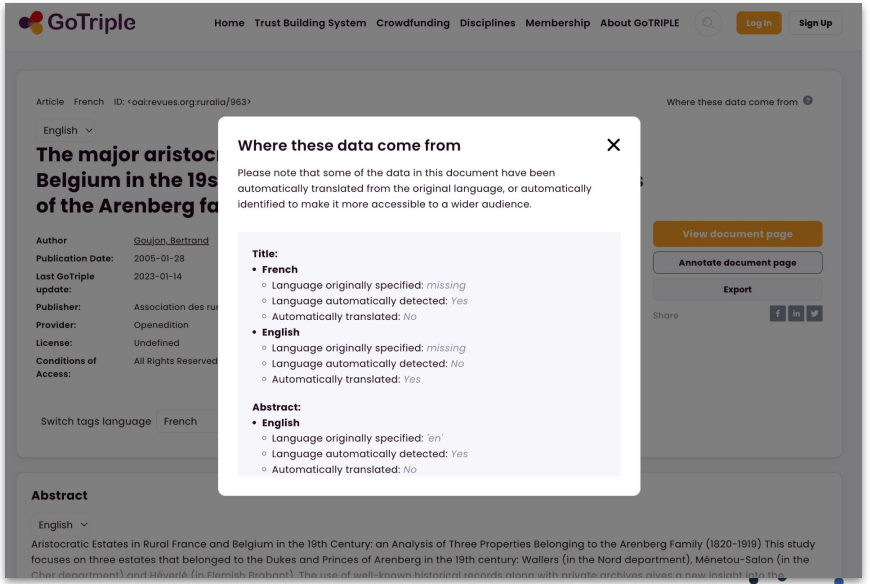
Document data enrichment

Multiple **new metadata** created by:

- **normalisation** rules
  - *controlled vocabularies* for languages, document types, licences, conditions of access
- ad-hoc **algorithms**
  - *clustering* to group duplicate documents
  - authors disambiguation
- **machine learning and NLP** techniques
  - classification and automatic annotation of documents and projects

Detailed presentation of the origin of data

Machine-actionable version easily available



net7

Document data enrichment

Multiple **new metadata** created by:

- **normalisation** rules
  - *controlled vocabularies* for languages, document types, licences, conditions of access
- ad-hoc **algorithms**
  - *clustering* to group duplicate documents
  - authors disambiguation
- **machine learning and NLP** techniques
  - classification and automatic annotation of documents and projects

Detailed presentation of the origin of data

Machine-actionable version easily available

The TRIPLE Ontologies

What's there:

- **multiple independent vocabularies** to **encourage reuse**
  - focus on the formalisation of **Document** entities
  - definition of the controlled vocabularies for document types, licenses, conditions of access, disciplines
  - document types linked to COAR entities
- new reusable **disciplines** vocabulary, **multilingual**, **linked** to multiple **standard ontologies** (Wikidata, LCSH, CESSDA, TRIPLE Vocabulary)

Coming next

- formalisation of Profile and linking with Document entities
- formalisation of Projects