Science and Technology Facilities Council

Good practices and guidelines for semantic interoperability

Alejandra Gonzalez-Beltran, PhD
Data and Software Engineering Group Leader

@alegonbel
@agbeltran
0000-0003-3499-8262
https://agbeltran.github.io/

http://doi.org/10.5281/zenodo.5559417

United Nations WORLD DATA FORUM

Multi-stakeholder Data Bridges II: making data work for cross-domain grand-challenges

4th October 2021

Why do we need semantic interoperability?

Semantic interoperability
is the ability of computer systems to exchange data with unambiguous, shared meaning

Why do we need semantic interoperability?

Global Challenges

→ Data-driven analyses

→ Standardised data collection

→ Data democratisation:
Collected data released often (following required privacy standards)

→ Open standards and machine-readable formats

Science and Technology Facilities Council

Why do we need semantic interoperability?

Global Challenges

→ Data-driven analyses

→ Standardised data collection

→ Data democratisation:
Collected data released often (following required privacy standards)

→ Open standards and machine-readable formats

→ Data integration challenges

↓

Need to resolve ambiguity of definitions from multiple data sources

↓

Semantic interoperability

Science and Technology Facilities Council

3 GOOD HEALTH AND WELL-BEING

Covid-19 Vaccination Open Data

Vaccines per age per area

area_name	area_code	date	dose	age_band	age_ higher	age_ lower	cum_ doses	new_ doses	population

Date	Country	Sex	AgeGroup	Population	Dose	NumberVaccinated	PercentCoverage	Cumulative NumberVaccinated	Cumulative PercentCoverage
numeric	text	text	text	numeric	text	numeric	numeric	numeric	numeric
Date on which vaccinnation occurred	7-digit country code	Sex of the individual	Age group of individual	Number of individuals in the cohort	2-dose schedule	Number of individuals who have received the vaccine	Percentage of cohort having received vaccination.	Cumulative number of individuals who have received the vaccine	Cumulative percentage of cohort having received vaccination

Data Dictionary

Daily vaccinations by age group and sex

Science and Technology Facilities Council

3 GOOD HEALTH AND WELL-BEING

Covid-19 Vaccination Open Data

Vaccines per age per area

area_name	area_code	date	dose	age_band	age_ higher	age_ lower	cum_ doses	new_ doses	population

Date	Country	Sex	AgeGroup	Population	Dose	NumberVaccinated	PercentCoverage	Cumulative NumberVaccinated	Cumulative PercentCoverage
numeric	text	text	text	numeric	text	numeric	numeric	numeric	numeric
Date on which vaccinnation occurred	7-digit country code	Sex of the individual	Age group of individual	Number of individuals in the cohort	2-dose schedule	Number of individuals who have received the vaccine	Percentage of cohort having received vaccination.	Cumulative number of individuals who have received the vaccine	Cumulative percentage of cohort having received vaccination

Data Dictionary

Daily vaccinations by age group and sex

Science and Technology Facilities Council

3 GOOD HEALTH AND WELL-BEING

Covid-19 Vaccination Open Data

Annotated with a term from a FAIR vocabulary

Vaccines per age per area

area_name	area_code	date	dose	age_band	age_ higher	age_ lower	cum_ doses	new_ doses	population

Date	Country	Sex	AgeGroup	Population	Dose	NumberVaccinated	PercentCoverage	Cumulative NumberVaccinated	Cumulative PercentCoverage
numeric	text	text	text	numeric	text	numeric	numeric	numeric	numeric
Date on which vaccinnation occurred	7-digit country code	Sex of the individual	Age group of individual	Number of individuals in the cohort	2-dose schedule	Number of individuals who have received the vaccine	Percentage of cohort having received vaccination.	Cumulative number of individuals who have received the vaccine	Cumulative percentage of cohort having received vaccination

Data Dictionary

Daily vaccinations by age group and sex

Values from FAIR vocabularies

Science and Technology Facilities Council

How do we achieve semantic interoperability?

… FAIR vocabularies

(vocabularies that adhere to the FAIR data principles)

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

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

Mark D. Wilkinson *et al.*#

https://doi.org/10.1038/sdata.2016.18

Science and Technology Facilities Council

What is a FAIR vocabulary?

Findable
indexed, listed

→ Registered in a community service or portal
Persistent identifiers for the vocabulary and terms

Accessible
on the web, downloadable

→ Reachable via a standard protocol
Machine/human readable representations

Interoperable
in a standard representation

→ Encoded in a standard representation language
Mappings and relationships to other vocabularies

Reusable
licensed, maintained

→ Clear and accessible license (ideally open) and with a maintenance process
Clear definitions and provenance metadata per term

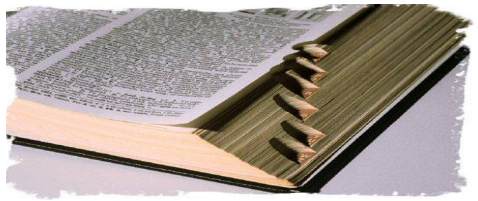
Why FAIR vocabularies?

- To know if references in different datasets mean the same thing, building data bridges
- To annotate data using terms in shared vocabularies to facilitate data integration
  - Classifications of places
  - Units of measure
  - Chemicals
  - Proteins
  - Taxa / species
  - ...
- Vocabularies trusted by the community, based on standards, machine actionable

How do you build a FAIR vocabulary?

PLOS COMPUTATIONAL BIOLOGY

Check for updates



RESEARCH ARTICLE

Ten simple rules for making a vocabulary FAIR

Simon J. D. Cox[1]*, Alejandra N. Gonzalez-Beltran[2], Barbara Magagna[3], Maria-Cristina Marinescu[4]

1 CSIRO Land and Water, Melbourne, Australia, **2** Science and Technology Facilities Council, Didcot, United Kingdom, **3** Environment Agency Austria, Wien, Austria, **4** Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain

* simon.cox@csiro.au

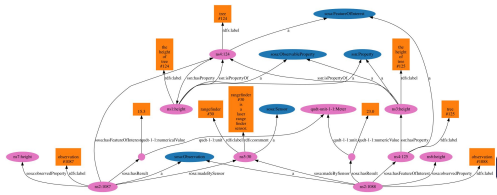
https://doi.org/10.1371/journal.pcbi.1009041

OPEN ACCESS

Citation: Cox SJD, Gonzalez-Beltran AN, Magagna B, Marinescu M-C (2021) Ten simple rules for making a vocabulary FAIR. PLoS Comput Biol 17(6): e1009041. https://doi.org/10.1371/journal.pcbi.1009041

Editor: Scott Markel, Dassault Systemes BIOVIA, UNITED STATES

Received: November 22, 2020



PLOS COMPUTATIONAL BIOLOGY

Check for updates

OPEN ACCESS

Citation: Cox SJD, Gonzalez-Beltran AN, Magagna B, Marinescu M-C (2021) Ten simple rules for making a vocabulary FAIR. PLoS Comput Biol 17(6): e1009041. https://doi.org/10.1371/journal.pcbi.1009041

Editor: Scott Markel, Dassault Systemes BIOVIA, UNITED STATES

Received: November 22, 2020

RESEARCH ARTICLE

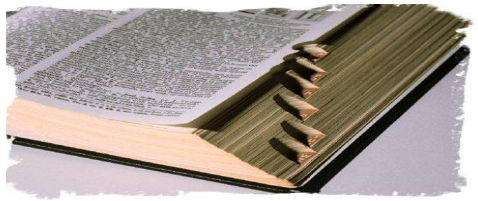
Ten simple rules for making a vocabulary FAIR

Simon J. D. Cox[1]*, Alejandra N. Gonzalez-Beltran[2], Barbara Magagna[3], Maria-Cristina Marinescu[4]

1 CSIRO Land and Water, Melbourne, Australia, **2** Science and Technology Facilities Council, Didcot, United Kingdom, **3** Environment Agency Austria, Wien, Austria, **4** Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain

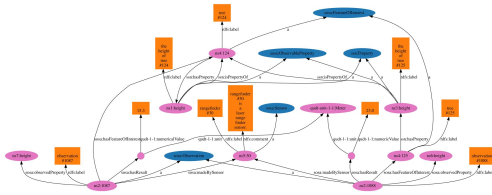
* simon.cox@csiro.au

https://doi.org/10.1371/journal.pcbi.1009041



Pre-conditions / assumptions

- community requirement to be able to **annotate or classify data**, using terms from a **controlled vocabulary**
- a legacy **vocabulary is available** (a list of terms and definitions)
- the vocabulary is *not* arranged & published to allow remote reference to individual terms e.g. **a print document, a digital document, or in a semi-structured form**
- **no other FAIR vocabulary** suitable for the application, and **acceptable to the community** is published



Ten simple rules for making a vocabulary FAIR

1. Determine the **governance** arrangements and custodian of the legacy vocabulary
2. Verify that the **legacy-vocabulary license** allows repurposing, and agree on the license for the **FAIR vocabulary**
3. Check **term** and **definition completeness** and **consistency** in the legacy vocabulary
4. Establish a **traceable maintenance-environment** for the FAIR vocabulary content
5. Assign a **unique and persistent identifier** to (a) the vocabulary and (b) each term in the vocabulary
6. Create **machine readable representations** of the vocabulary terms
7. Add vocabulary **metadata**
8. **Register** the vocabulary
9. Make the vocabulary **accessible** for humans and machines
10. Implement a **process** for publishing **revisions** of the FAIR vocabulary

FAIR Vocabularies

A website for the guidelines for FAIR vocabularies.

Guidelines

Web Standards

Make vocabulary FAIR

Examples

About

View Organization on Github

Cross-discipline data discovery, integration, and synthesis remain challenging tasks for technical, social and content-related reasons. However, those tasks are crucial for addressing global scientific and societal challenges. Understanding the data, identifying the terminology used to annotate them, and how they relate is a prerequisite to enable data integration.

Here we provide a series of guidelines that help on creating and using **FAIR** (Findable, Accessible, Interoperable & Reusable) vocabularies.

What is a FAIR vocabulary?

The FAIR guiding principles for scientific data management and stewardshipt provide recommendations to make data Findable, Accessible, Interoperable and Reusable.

One of the principles indicate that to be interoperable:

> *I2. (meta)data use vocabularies that follow FAIR principles*

In this project we provide guidance for FAIR vocabularies, which are crucial to producing FAIR data.

For a vocabulary to be FAIR, it should be:

- **Findable**: be registered (indexed, listed) in a vocabulary service
- **Accessible**: be available on the web, downloadable
- **Interoperable**: encoded in a standard representation, such as the Web Ontology Language (OWL) or SKOS and domain-specific extensions
- **Reusable**: licensed and maintained, ideally with an open license

FAIR vocabularies representation relies on web standards

Guidelines for FAIR Vocabularies

- Guidelines for making a vocabulary FAIR
- Guidelines for maintaining a FAIR vocabulary - coming soon!

https://fairvocabularies.github.io

- A growing number of resources and guidelines supporting the lifecycle of FAIR vocabularies
- Guidelines, examples, pointers to technical best practices
  - How to find existing vocabularies
  - How to maintain them
  - ....

Science and Technology Facilities Council

Take-away messages

- **Data-driven analyses and solutions** are crucial for addressing global challenges in an **evidenced-based approach**
- These approaches pose many data integration **challenges**
- **Semantic interoperability** is an **enabler** for **integrating data**, as it resolves the ambiguity in term definitions
- **FAIR vocabularies** are needed for semantic interoperability and can be used as data bridges
- More **guidelines**, **tools**, and **training** are needed to facilitate the use of open standards and machine-readable formats for FAIR vocabularies and more broadly, data sharing

Science and Technology Facilities Council

Questions?

Science and Technology Facilities Council

Thank you

Science and Technology Facilities Council @STFC_Matters Science and Technology Facilities Council