FAIR
OFFICE AUSTRIA

FAIR Principles

Samah Jaber, FAIR Office Austria, 13th Sep 2022

with slides from lecture" FAIR Principles" by Tomasz Miksa

TECHNISCHE UNIVERSITÄT WIEN

TU Graz

universität wien

This PPT is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).
It is attributed to Samah Jaber (with contribution of slides by Tomasz Miksa). June 2022

About the Speaker

Samah Jaber BSc, MSc

✓*Areas of Specialties*

- IT Governance & Operational Management
- Quality Management System
- Process Compliance/Improvement
- Information Architecture.
- FAIRification Process

MSc. in Data Science at TUWIEN

✓*Research Interests:*

- Data Stewardship
- Data Management Plan
- FAIR Data

AGENDA

- Why we need FAIR principles?
- FAIR principles (big picture)
- FAIR principles (details)
- How repositories support FAIRness?
- How to FAIRify data?
- FAIR assessment
- GO FAIR
- FAIR Office Austria

Why we need FAIR Principles?

Researchers trying to reuse data…

Conversation of two researchers

- Can I see your data?
- It's on my USB stick
- Can I have it?
- I have in a box and I have moved recently
- Can I have it?
- I forgot to label the boxes…
- (half a year later)
- Thanks, for the USB. However, I cannot read the hexadecimal file on it. How do I open it?
- You need a special program
- What program?
- …

Hanson, Karen; Surkis, Alisa; Yacobucci, Karen: Data Sharing and Management Snafu in 3 Short Acts. https://doi.org/10.5446/31036

FAIR
OFFICE AUSTRIA

Data

- The output of any research processing (e.g: capturing, pre-processing, transformation, integration, analysis).

Search

- To process data we need to make them searchable.
- We're producing more and more data all the time, and if we don't add structure to that data, we only make it useful for ourselves.

FAIR

- Provide a way to extend the reach of our data.
- Tend to allow more people to make use of it.
- Enhance your recognition as a researcher.

Research Community

- **Researchers** FAIR data will result in better research output.
- **Funders** FAIR data will add more value to publicly funded research.
- **Publishers** FAIR data will improve the review process.
- **Universities** FAIR data will ensure high research integrity.

FAIR Principles and Research Lifecycle?

These four principles should be applied throughout the entire data lifecycle, and they are closely interconnected to increase the reuse of scientific data.

www.nature.com/scientificdata

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.*#

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

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.

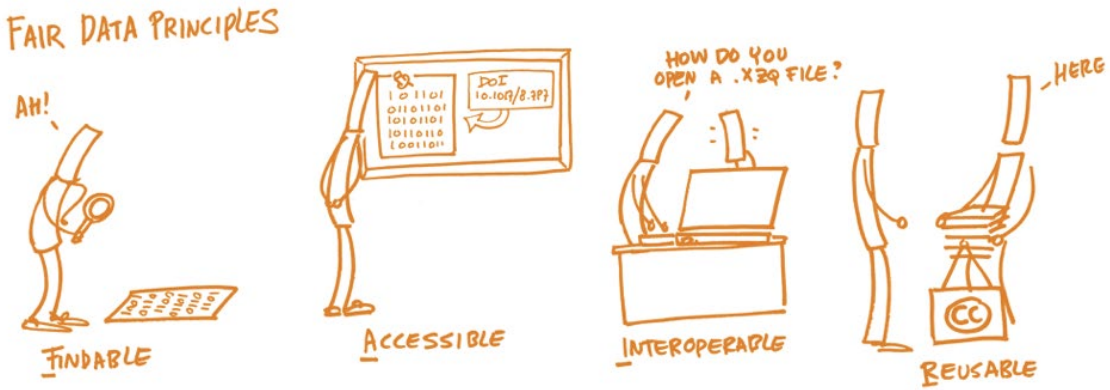
Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects[2]—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

Correspondence and requests for materials should be addressed to B.M. (email: barend.mons@dtls.nl).
#A full list of authors and their affiliations appears at the end of the paper.



https://www.nature.com/articles/sdata201618

https://www.go-fair.org/fair-principles/

FAIR OFFICE AUSTRIA

FAIR vs fair

FAIR principles =/= Algorithmic fairness

To be FAIR

- To apply/use FAIR principles
- Focus on how data is managed, etc.

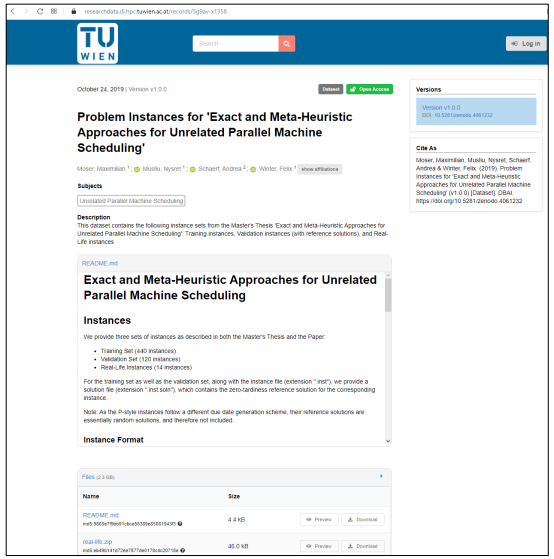
To be fair

- Evade bias
- Focus on design and implementation

FAIR Principles (big picture)

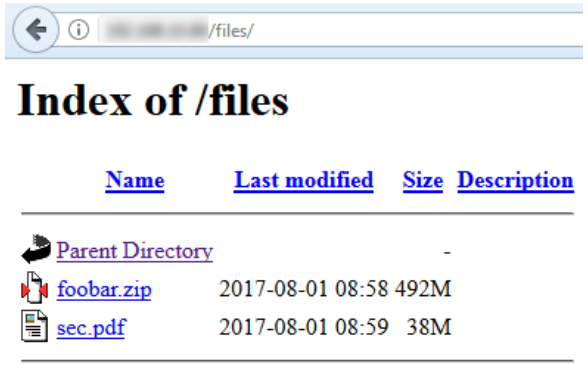
Findable – simplified examples

- Yes



Data repository

- No

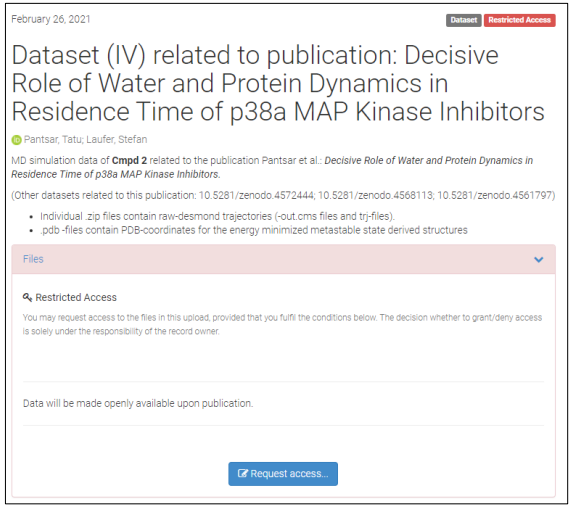


Personal website

Accessible – simplified examples

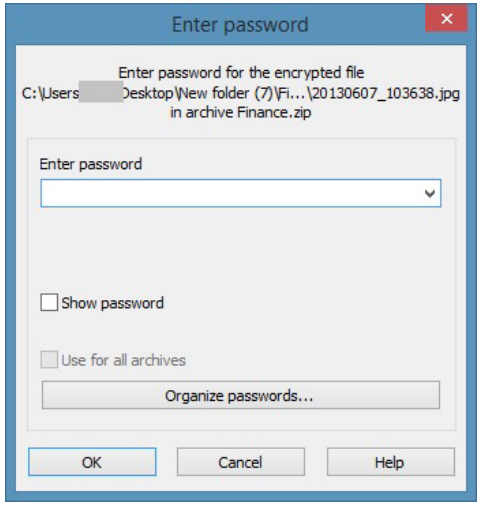
FAIR OFFICE AUSTRIA

- Yes



Restricted access, but a clear way to request access

- No



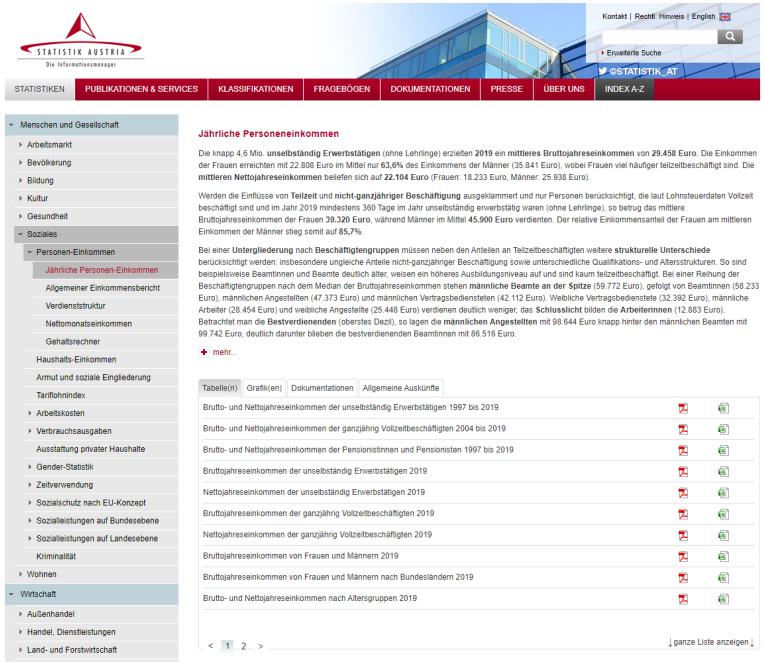
Interoperable – simplified examples

FAIR OFFICE AUSTRIA

- Yes
  - XML following known XSD Schema
  - MP3 for audio recordings
- No
  - Custom XML without any documentation
  - M4P (Apple) for audio recordings

Reusable – simplified examples

- Yes



Trusted source, permission to reuse, well defined meaning of terms used

- No

Provenance and permissions not clear

Metadata

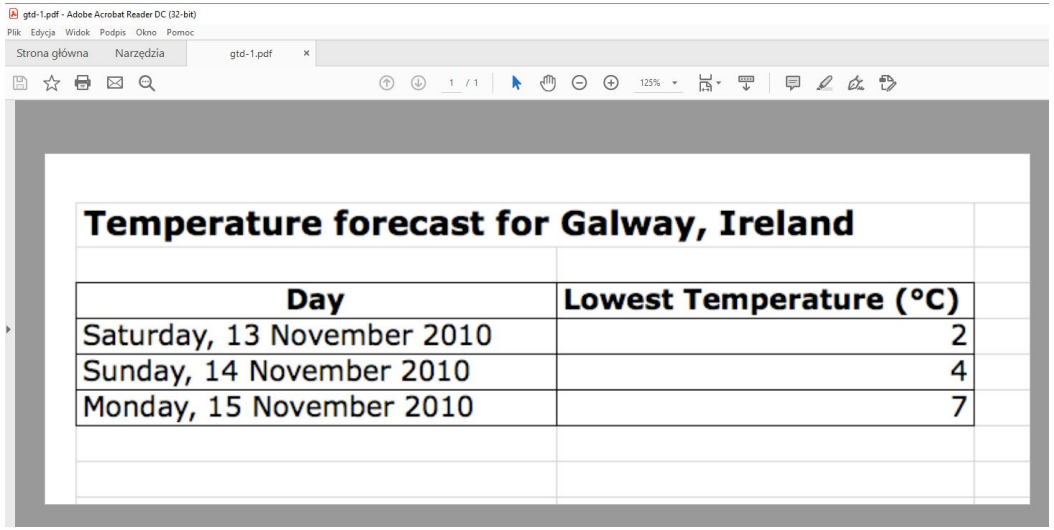
- What is in the picture?
- Where has it been produced?
- Who has produced it?
- What are the ingredients ?
- Production date?
- Valid till?

FAIR vs Open Data

- FAIR data =/= open data!

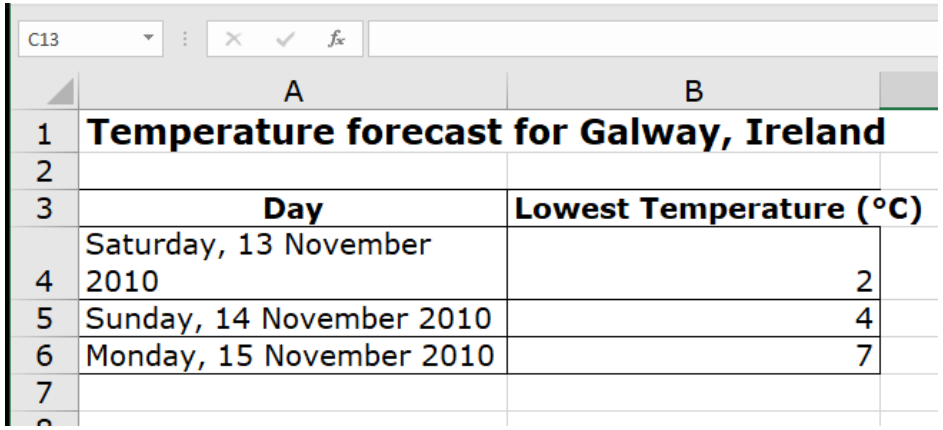
Even if the data itself cannot be shared openly, you should create and publish a description of your data so that researchers with a relevant purpose can request permission to reuse the data.

FAIR and Machine-actionability



Not machine-actionable

Machine-actionable



FAIR Principles (details)

Findable

- F1. (Meta)data are assigned a globally unique and **persistent identifier**
- F2. Data are described with rich **metadata**
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a **searchable resource**

Persistent identifiers (F1)

- Example
  - A car has only one VIN (PID), but can have many number plates over its lifetime (URL)

1

0..*

Persistent Identifiers (F1)

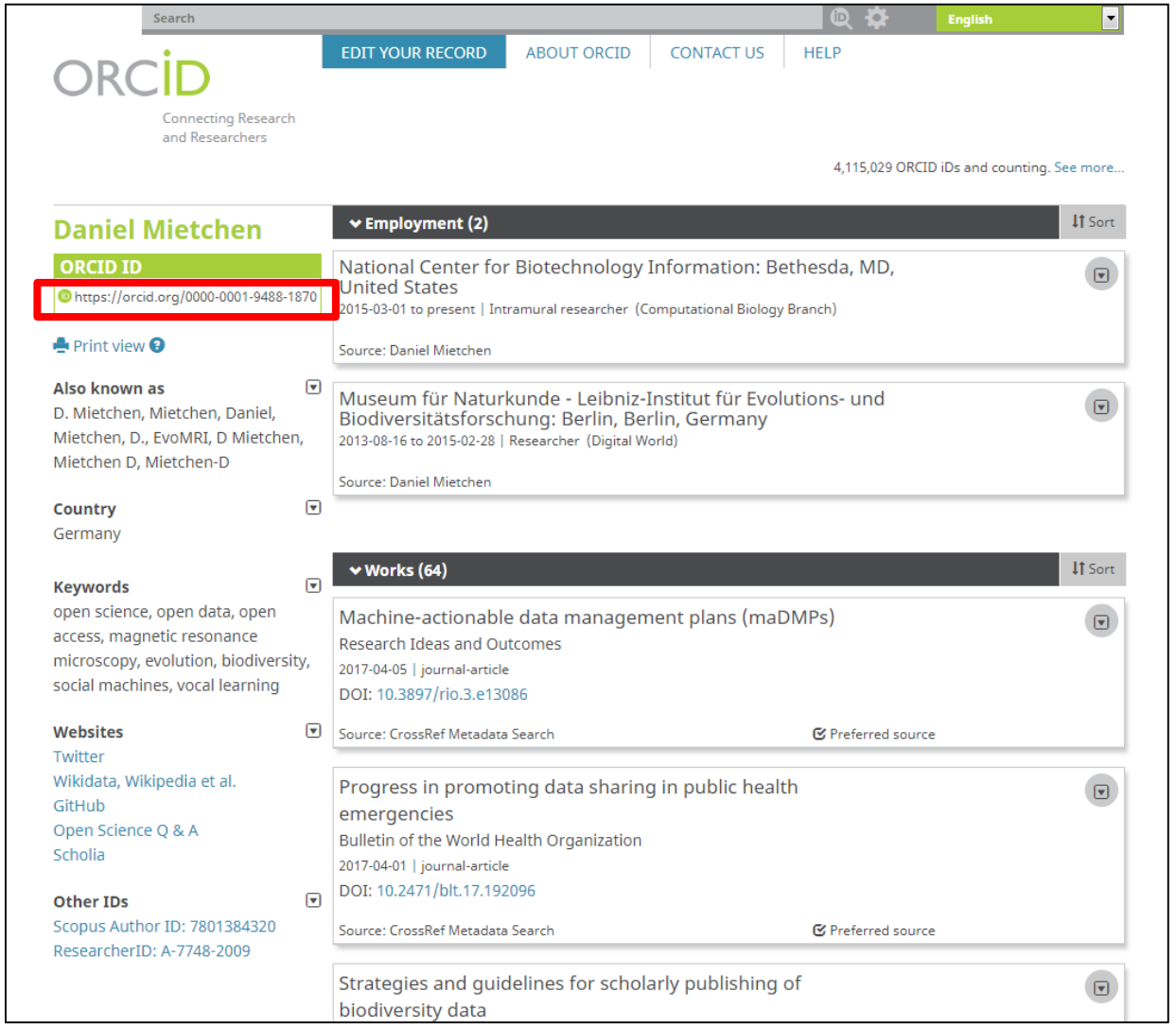
- Digital Object Identifier (DOI)
  - Uniquely identify objects
  - DOI assigned once
  - Physical location of data can change

10.5281/zenodo.1068223

- ORCID ID
  - Unique person ID
  - ORCID assigned once
  - Person can change affiliations (jobs)

0000-0002-4929-7875

ORCID Example



DOI example – assigned to publication

PLOS COMPUTATIONAL BIOLOGY

EDUCATION

Ten principles for machine-actionable data management plans

Tomasz Miksa[1☯*], **Stephanie Simms**[2☯], **Daniel Mietchen**[3☯], **Sarah Jones**[4☯]

1 SBA Research & TU Wien, Vienna, Austria, **2** California Digital Library, University of California, Oakland, United States of America, **3** Data Science Institute, University of Virginia, Charlottesville, United States of America, **4** Digital Curation Centre, Glasgow, United Kingdom

☯ These authors contributed equally to this work.
* miksa@ifs.tuwien.ac.at

Abstract

Data management plans (DMPs) are documents accompanying research proposals and project outputs. DMPs are created as free-form text and describe the data and tools employed in scientific investigations. They are often seen as an administrative exercise and not as an integral part of research practice.

There is now widespread recognition that the DMP can have more thematic, machine-actionable richness with added value for all stakeholders: researchers, funders, repository managers, research administrators, data librarians, and others. The research community is moving toward a shared goal of making DMPs machine-actionable to improve the experience for all involved by exchanging information across research tools and systems and embedding DMPs in existing workflows. This will enable parts of the DMP to be automatically generated and shared, thus reducing administrative burdens and improving the quality of information within a DMP.

This paper presents 10 principles to put machine-actionable DMPs (maDMPs) into practice and realize their benefits. The principles contain specific actions that various stakeholders are already undertaking or should undertake in order to work together across research communities to achieve the larger aims of the principles themselves. We describe existing initiatives to highlight how much progress has already been made toward achieving the goals of maDMPs as well as a call to action for those who wish to get involved.

Check for updates

OPEN ACCESS

Citation: Miksa T, Simms S, Mietchen D, Jones S (2019) Ten principles for machine-actionable data management plans. PLoS Comput Biol 15(3): e1006750. https://doi.org/10.1371/journal.pcbi.1006750

CANADA

Published: March 28, 2019

Copyright: © 2019 Miksa et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

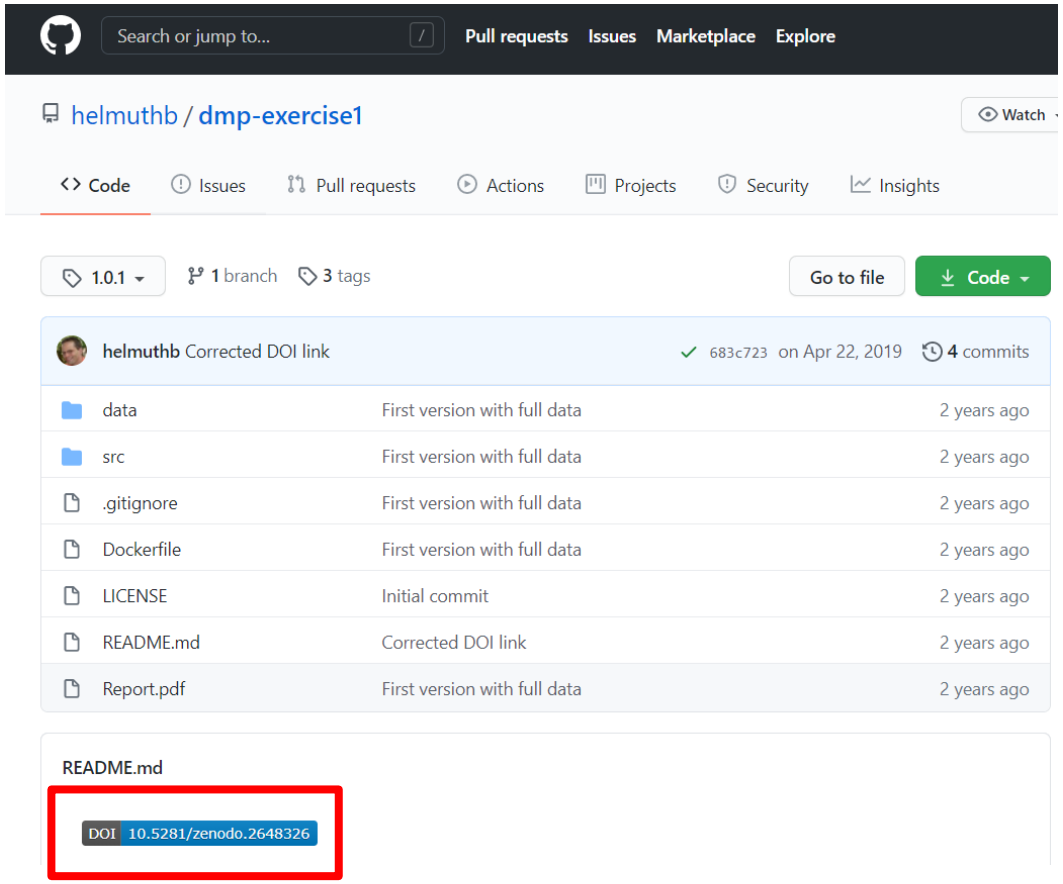
Funding: This research was carried out in the context of the Austrian COMET K1 program and publicly funded by the Austrian Research Promotion Agency (FFG) and the Vienna Business Agency (WAW). It was also supported by an NSF EAGER grant awarded to the California Digital Library (Award Number 1745675). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

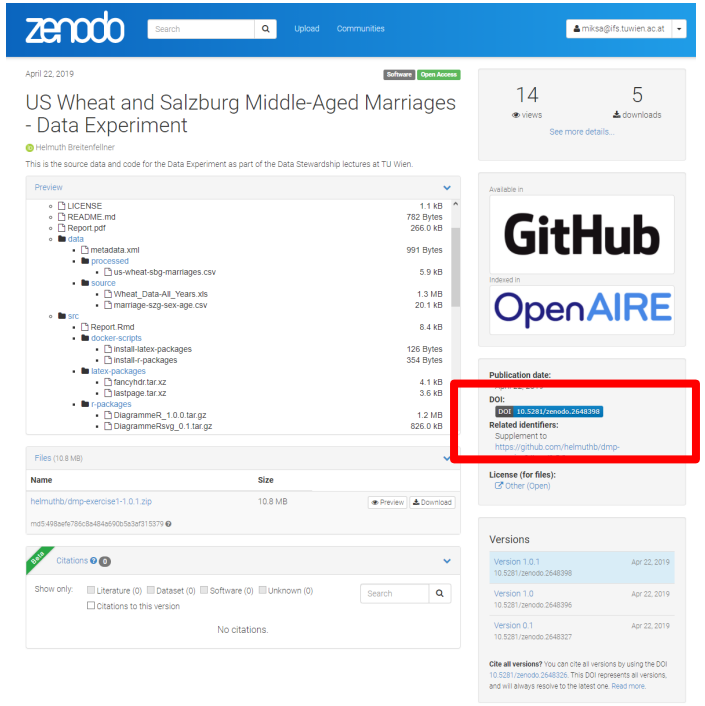
Competing interests: The authors have declared that no competing interests exist.

Introduction

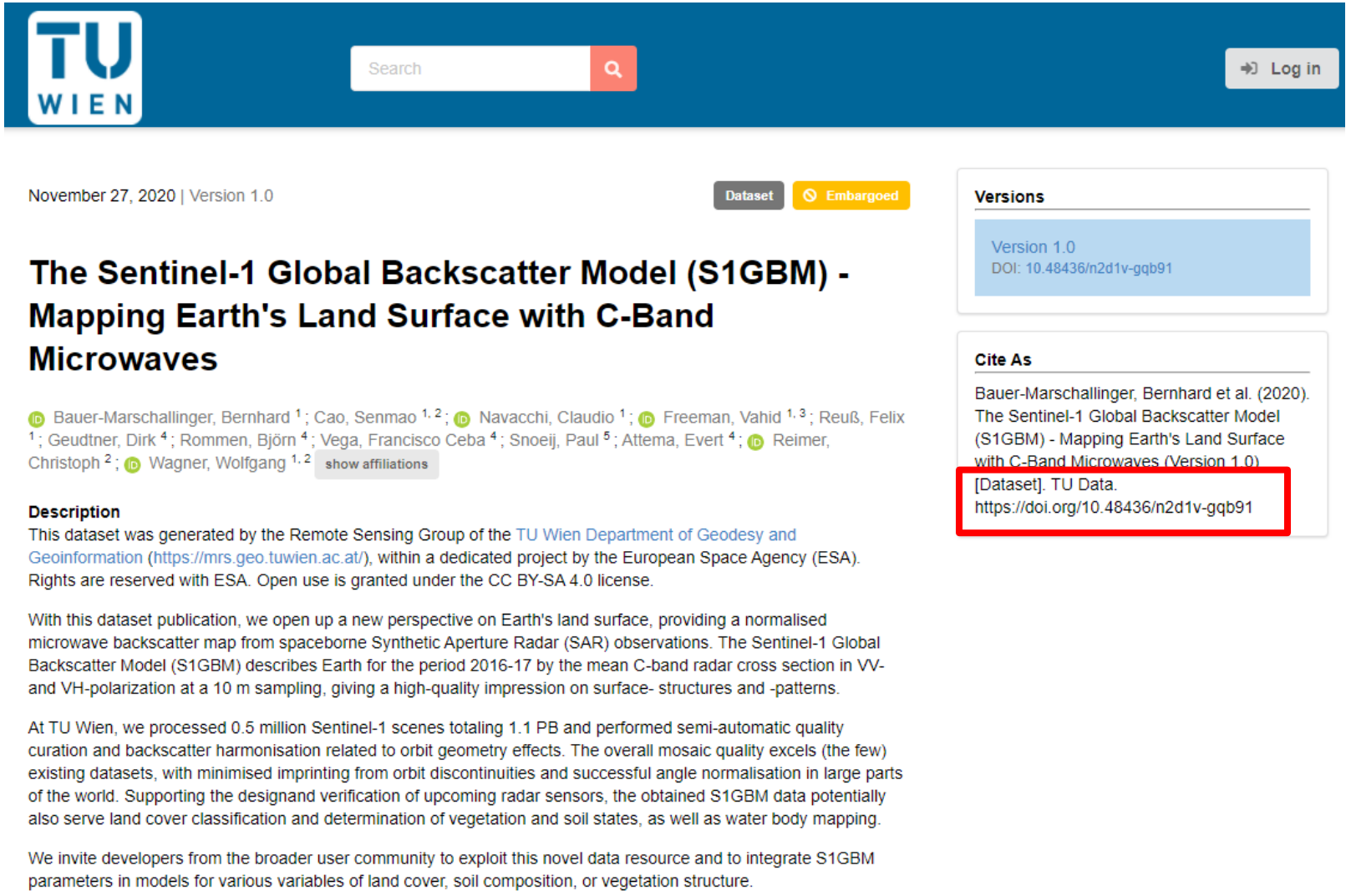
Data management plans (DMPs) are documents accompanying research proposals. They describe the data that are used and produced during the course of research activities, where the data will be archived, which licenses and constraints apply, and to whom credit should be given. DMPs are awareness tools to help researchers manage their data and ensure that it will be of high quality, accessible, and reusable after the project has ended. DMPs are typically created manually, mostly by researchers using checklists and online questionnaires. They are required by funding bodies and institutions all over the world, e.g., the National Science

DOI example – assigned to code





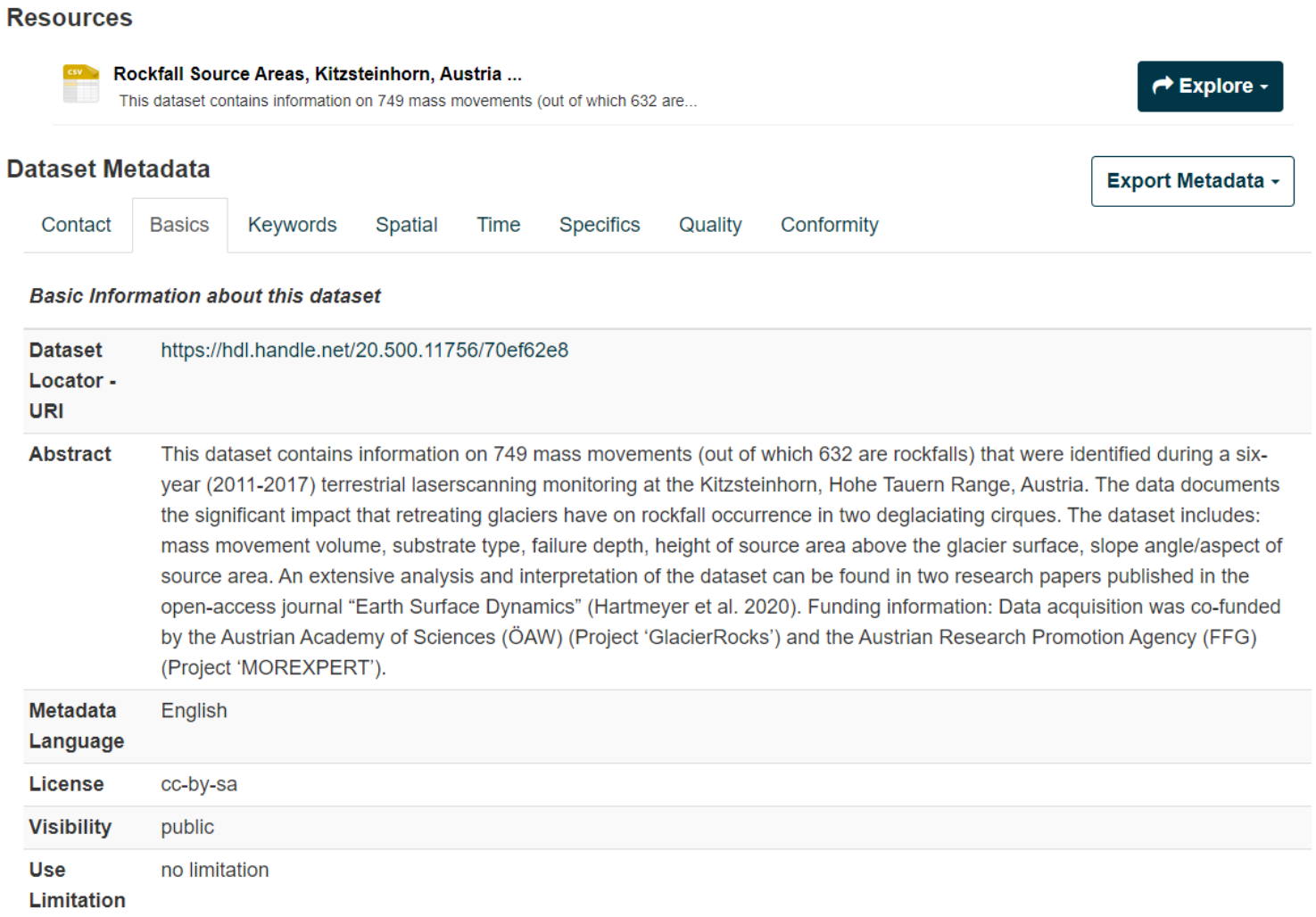
DOI example - assigned to data



PIDs Examples

F2. Data are described with rich metadata

FAIR OFFICE AUSTRIA



Resources

Rockfall Source Areas, Kitzsteinhorn, Austria ...
This dataset contains information on 749 mass movements (out of which 632 are...

Explore

Dataset Metadata

Export Metadata

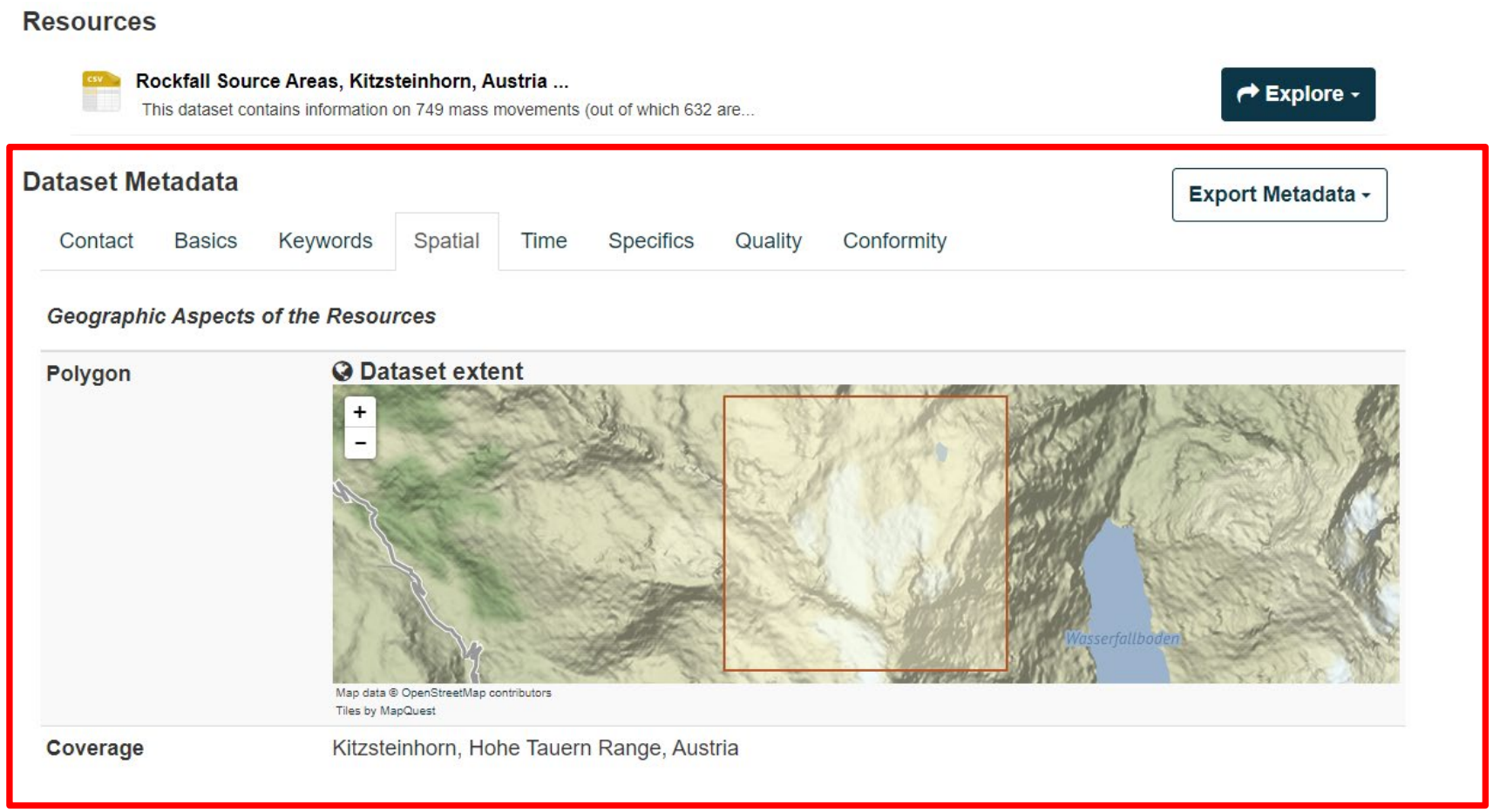
Contact | Basics | Keywords | Spatial | Time | Specifics | Quality | Conformity

Basic Information about this dataset

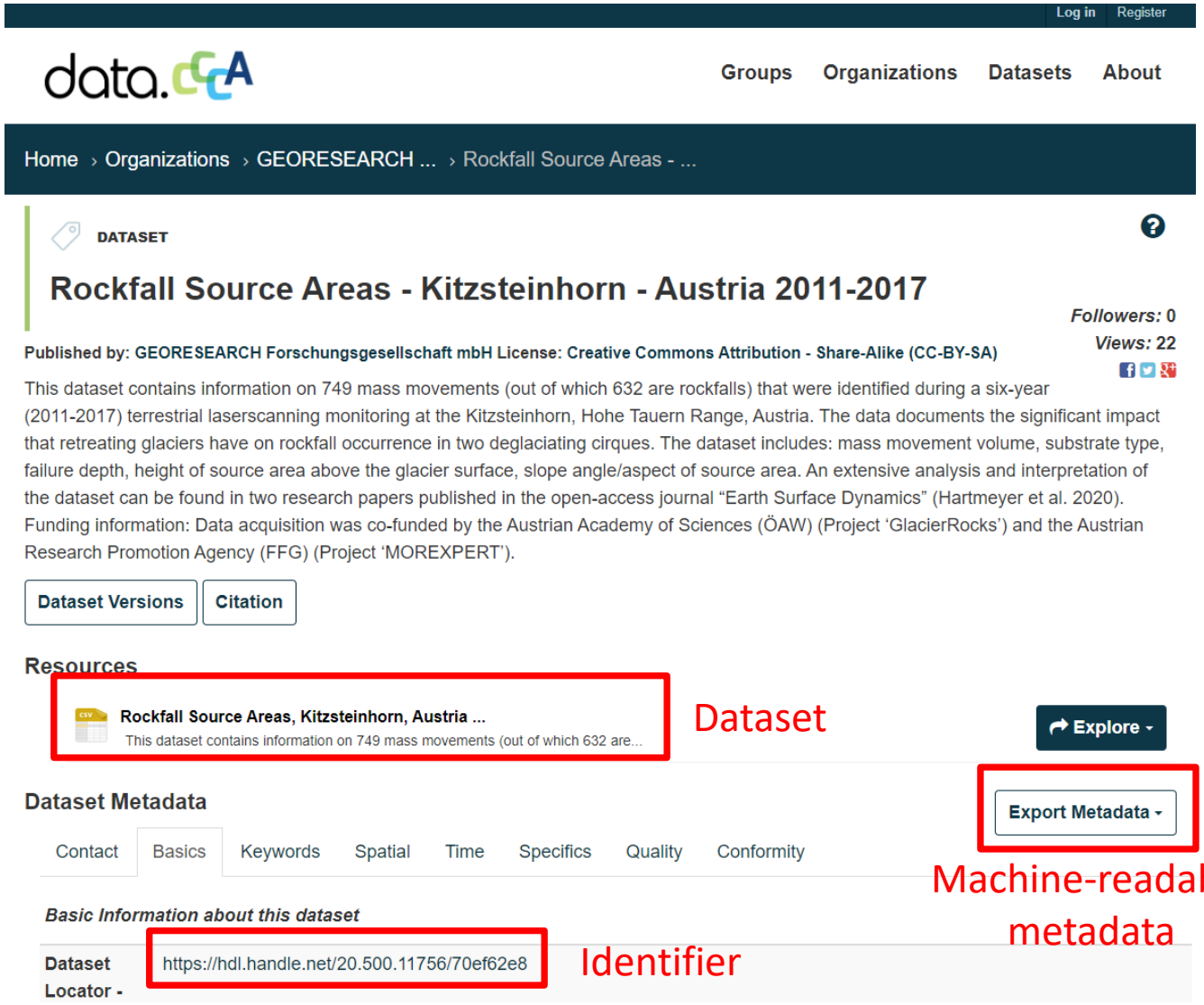
Dataset Locator - URI	https://hdl.handle.net/20.500.11756/70ef62e8
Abstract	This dataset contains information on 749 mass movements (out of which 632 are rockfalls) that were identified during a six-year (2011-2017) terrestrial laserscanning monitoring at the Kitzsteinhorn, Hohe Tauern Range, Austria. The data documents the significant impact that retreating glaciers have on rockfall occurrence in two deglaciating cirques. The dataset includes: mass movement volume, substrate type, failure depth, height of source area above the glacier surface, slope angle/aspect of source area. An extensive analysis and interpretation of the dataset can be found in two research papers published in the open-access journal "Earth Surface Dynamics" (Hartmeyer et al. 2020). Funding information: Data acquisition was co-funded by the Austrian Academy of Sciences (ÖAW) (Project 'GlacierRocks') and the Austrian Research Promotion Agency (FFG) (Project 'MOREXPERT').
Metadata Language	English
License	cc-by-sa
Visibility	public
Use Limitation	no limitation

F2. Data are described with rich metadata

FAIR OFFICE AUSTRIA

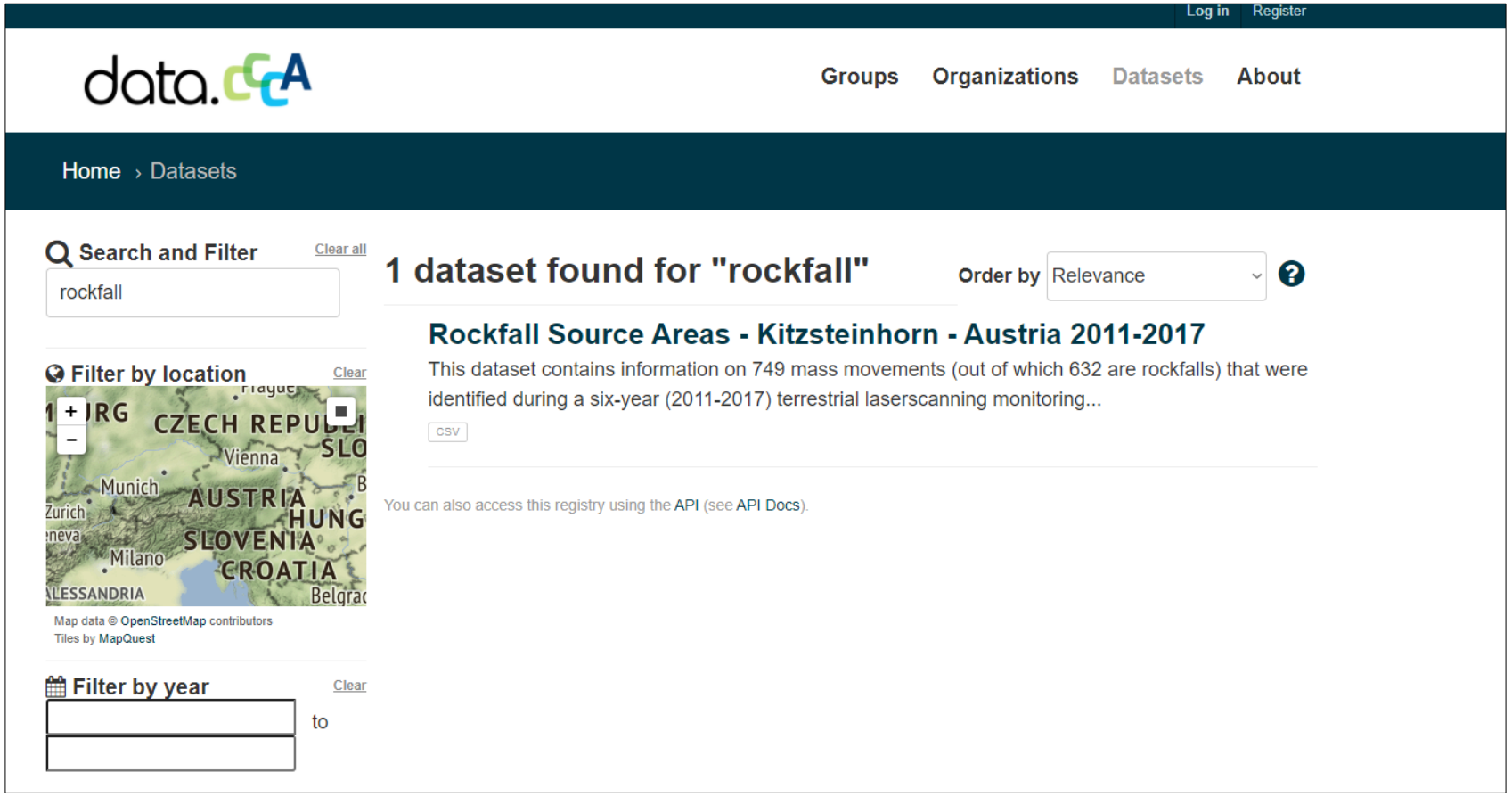


F3. Metadata clearly and explicitly include the identifier of the data they describe



https://hdl.handle.net/20.500.11756/70ef62e8

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



https://data.ccca.ac.at

Accessible

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
  - A1.1 The protocol is open, free, 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

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

FAIR OFFICE AUSTRIA

- Retrieve data by ‘clicking on a link’, without specialized tools or communication methods (http, ftp).
- “Anyone with a computer and an internet connection can access at least the metadata”
- HTTP
  - Open – specification of the protocol is known to everyone
  - Free – no need to pay to “use Internet”
- FAIR data ≠ Open data.

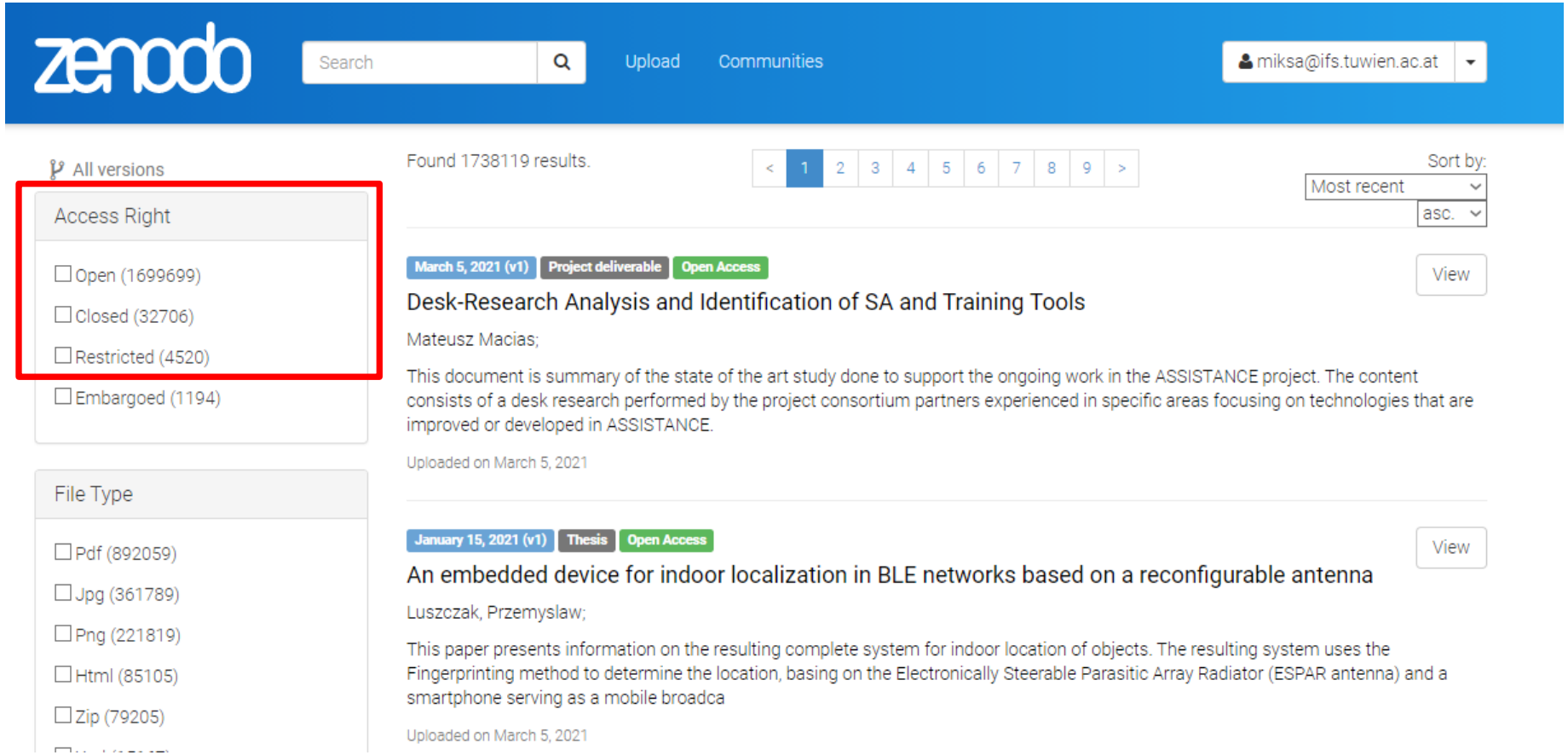
OSI model

Layer	Name	Example protocols
7	Application Layer	HTTP, FTP, DNS, SNMP, Telnet
6	Presentation Layer	SSL, TLS
5	Session Layer	NetBIOS, PPTP
4	Transport Layer	TCP, UDP
3	Network Layer	IP, ARP, ICMP, IPSec
2	Data Link Layer	PPP, ATM, Ethernet
1	Physical Layer	Ethernet, USB, Bluetooth, IEEE802.11

A1.2 protocol allows for authentication and authorization

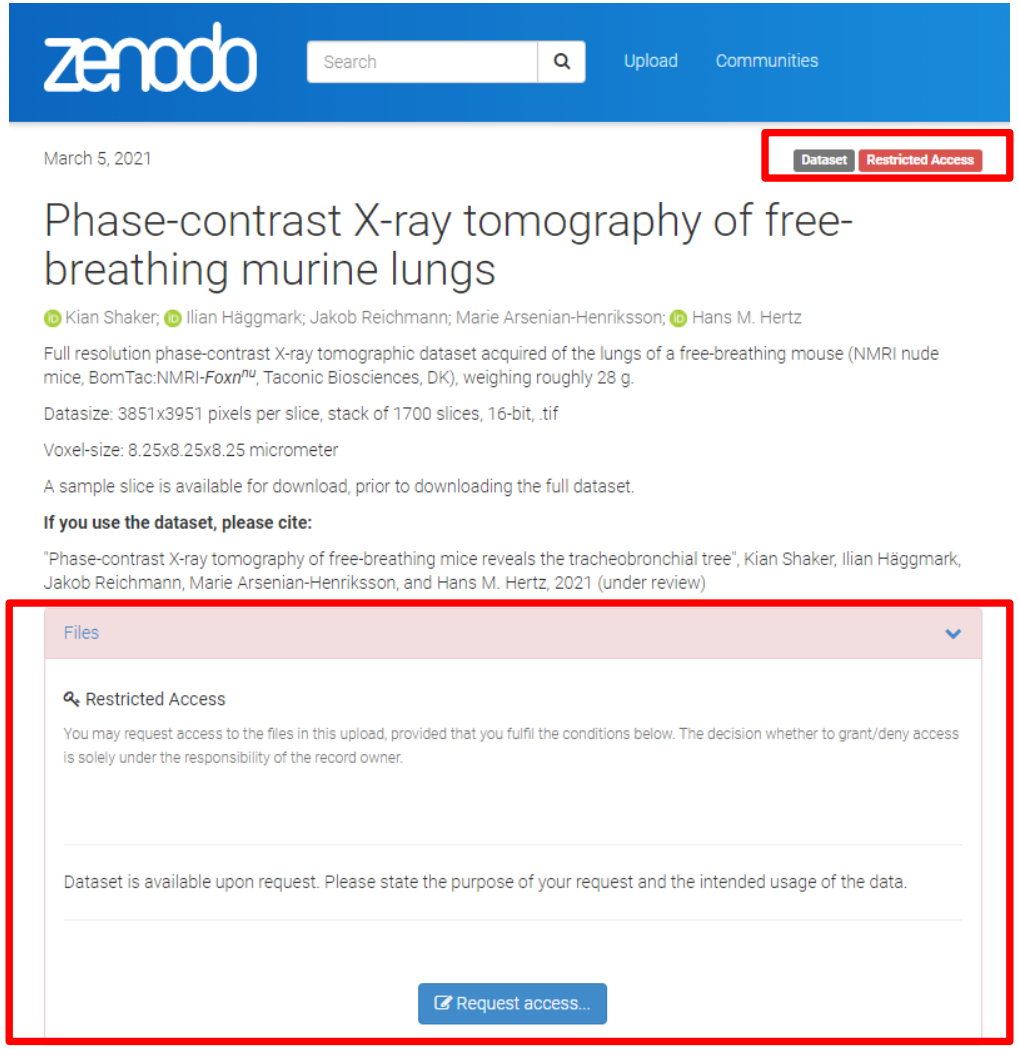
- Protected and private data can be FAIR
- Possible types of access
  - Open – everyone has access
  - Shared or restricted – only a selected/ invited group of people can access
  - Closed – only the owner has access

Accessible - example



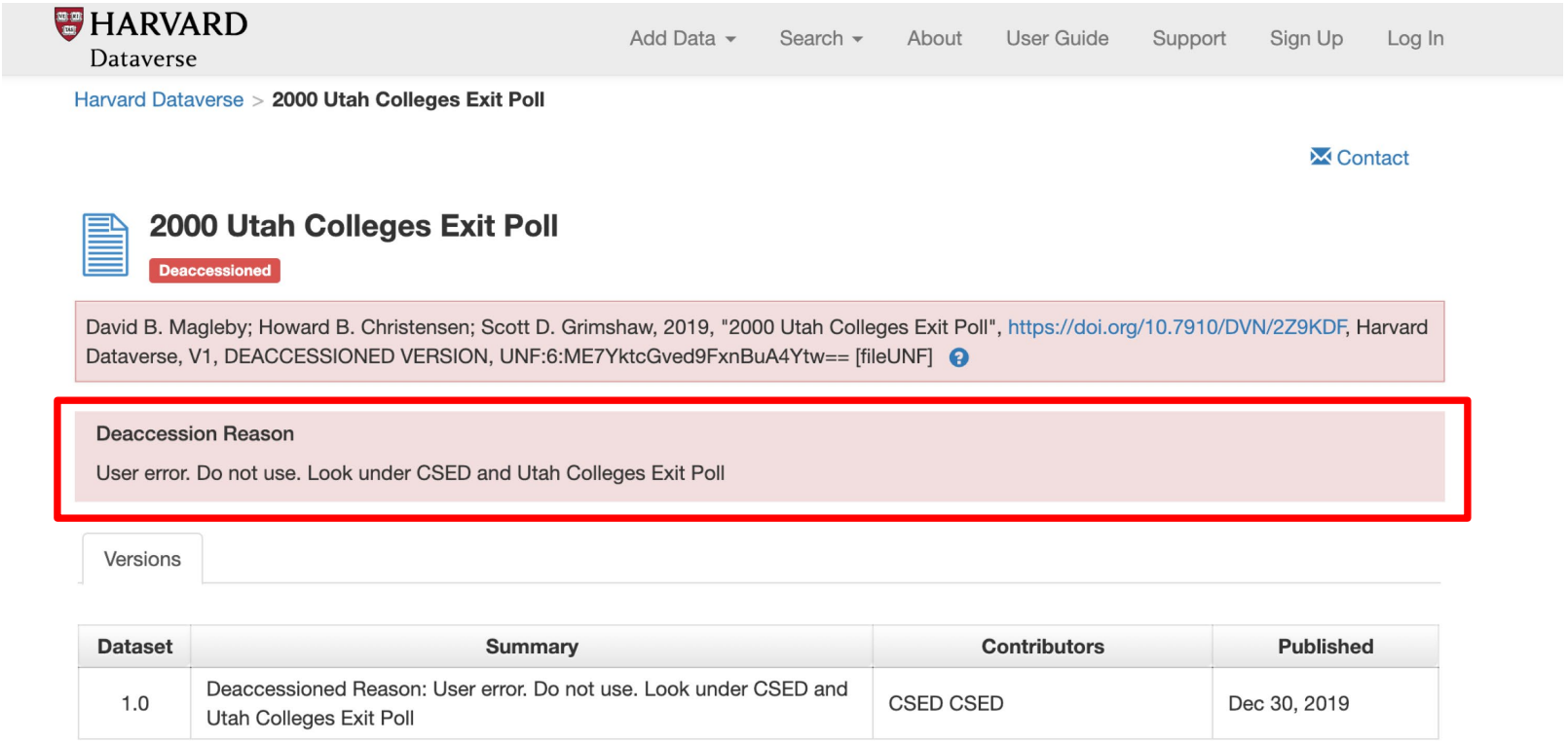
https://zenodo.org/search?page=1&size=20&q=

Accessible - example



Tombstone pages (A2)

- Metadata is accessible, even when the data is no longer available



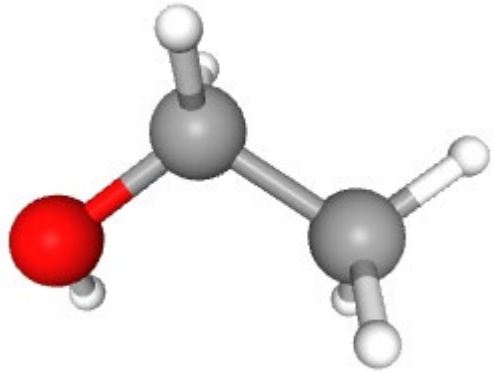
Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable **language for knowledge representation**.
- I2. (Meta)data use **vocabularies** that follow FAIR principles
- I3. (Meta)data include **qualified references** to other (meta)data

Language for knowledge representation (I1)

FAIR OFFICE AUSTRIA

- “Data that should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings”
- Use:
  - Common formats
    - RDF, JSON (+schema),
    - CSV (+ good README)
  - Well defined/described data models
  - Known representations
    - e.g. InChi Key: IDGUHHHQCWSQLU-UHFFFAOYSA-N
    - Data to have a standard way to describe itself to whomever (machine or human) try to access it without a need for a translator or mappings (common Format: JSON, RDF…)



Vocabularies

- Vocabulary: Computer-readable file that define meta(data).
- Help evade ambiguities
- “My plane lands in London…” – where exactly?

County	ICAO	IATA	Airport name	Usage
Greater London	EGKB	BQH	*London Biggin Hill Airport*	Public
Greater London	EGML		Damyns Hall Aerodrome	Private
Greater London	EGLL	LHR	***Heathrow Airport***	Public
Greater London	EGWU	NHT	*RAF Northolt*	Military
Greater London	EGLC	LCY	***London City Airport***	Public
Greater London	EGLW		*London Heliport*	Public

- Controlled vocabularies: IATA and ICAO

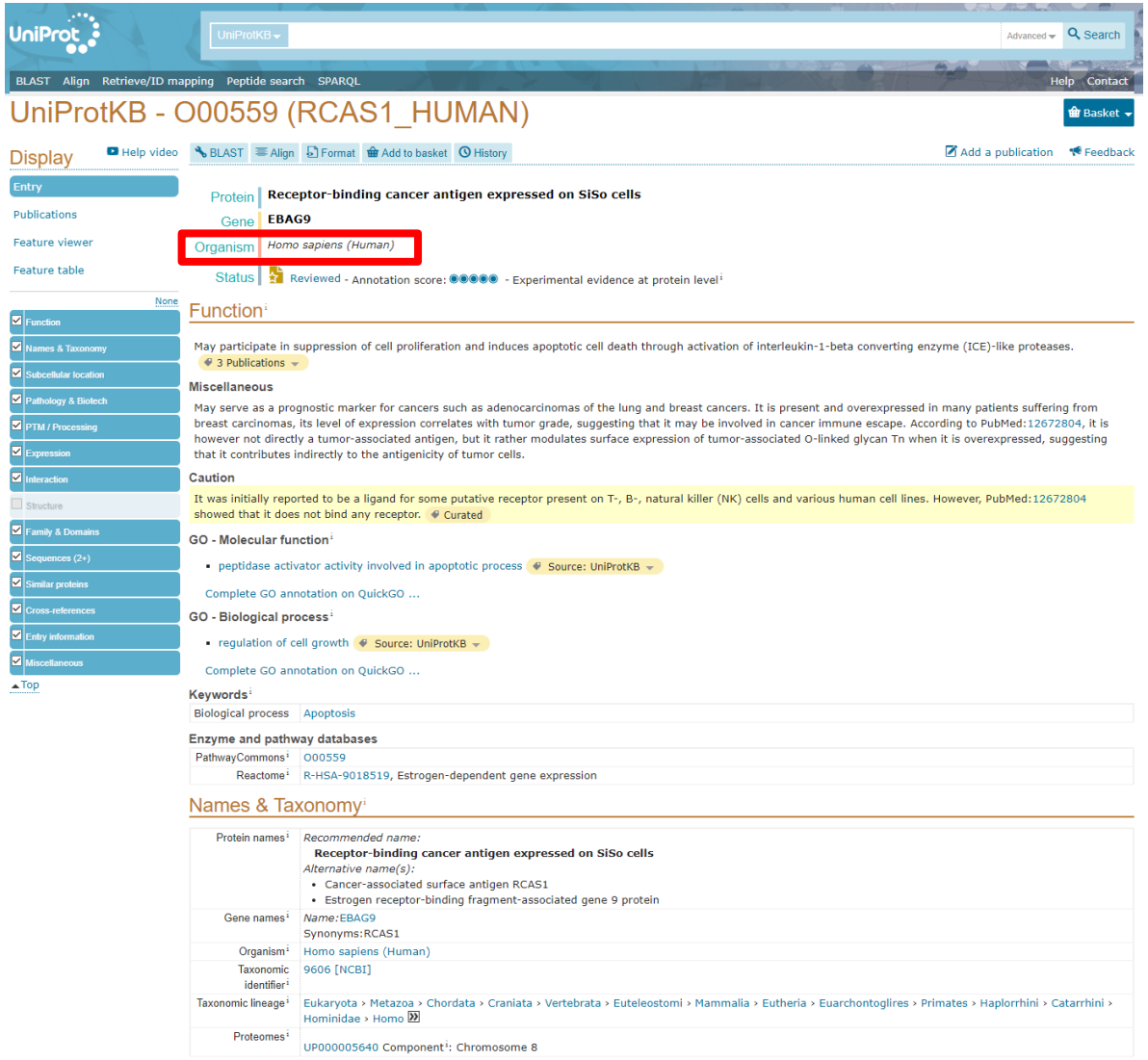
https://en.wikipedia.org/wiki/List_of_airports_in_the_United_Kingdom_and_the_British_Crown_Dependencies

Vocabularies

- Less time/money spent on data cleaning
  - Different languages
  - Spelling mistakes
  - Abbreviations
  - Capital letters

Vienna	**Beč** (Croatian, Serbian, older Bulgarian), **Beç** (older Turkish)*, **Bech** or **Vidnya** (Romani), **Bécs** (Hungarian)*, **Bin** / **Pin** - 빈 (Korean), **Dunaj** (Slovene)*, **Fienna** (Welsh), **Vedunia** (Celtic), **Vena** - Вена (Russian), **Vídeň** (Czech)*, **Viden'** / **Videň** (Ukrainian)*, **Viedeň** (Slovak), **Viên** (Vietnamese), **Viena** / **Vijena**/ **Виена** (Belarusian, Bulgarian, Macedonian), **Viena** (Catalan*, Lithuanian, Portuguese*, Romanian*, Spanish*, Tagalog*), **Vienna** (Italian)*, **Vienne** (French)*, **Viénni** - *Βιέννη* (Greek), **Vieno** (Esperanto), **Viin** (Estonian), **Vin** - ווין (Yiddish), **Vín** (Irish, Icelandic), **Vina** - וינה (Hebrew), **Vínarborg** (Icelandic variant), **Vindobona** (Latin), **Vīne** (Latvian)*, **Viyana** (Turkish)*, **Vjenë** (Albanian), **Vjenna** (Maltese), **Vyana** (Azeri), **Wean** (local Viennese, Austrian and Bavarian dialects)*, **Weiyena** - 維也納 (Chinese)*, **Wene** (Afrikaans), **Wenen** (Dutch)*, **Wiedeń** (Polish)*, **Wien** (Danish*, Finnish*, German*, Norwegian*, Swedish*), **Wīn** - ウィーン (Japanese)*, **Wina** (Indonesian), **فيينا** (Arabic), **وین** (Persian)

Vocabularies (I2)



https://www.uniprot.org/uniprot/O00559

Vocabularies (I2)

- Each metadata field has its definition

Organism

Last modified April 10, 2018

This subsection of the Names and taxonomy section provides information on the name(s) of the organism that is the source of the protein sequence.

The organism designation consists of the Latin scientific name, usually composed of the genus and species names (the binomial system developed by Linnaeus), followed optionally by the English common name and a synonym.
Examples: Bacillus subtilis, Homo sapiens (Human), Cardamine pratensis (Cuckoo flower) (Alpine bitter cress)

The synonym can be a common name in English (or in Latin in the case of some historical legacy names).
Example: Radianthus magnifica (Magnificent sea anemone) (Heteractis magnifica)

In the case of viruses, the designation does not follow the binomial system. The English common name is used as the scientific name, sometimes followed by an acronym. When possible, viruses are named according to the nomenclature of the International Committee on Taxonomy of Viruses (ICTV).
Examples: Human immunodeficiency virus type 1 (isolate BRU/LAI group M subtype B) (HIV-1) , Influenza A virus (strain A/Aichi/2/1968 H3N2)

The organism name can differ from that given by the international nucleotide sequence databases for the same taxon. This is mainly due to our efforts in providing the most descriptive common names and synonyms to our users.

Note that the proteome for a given organism, when available, can be accessed through the proteomes page of our website.

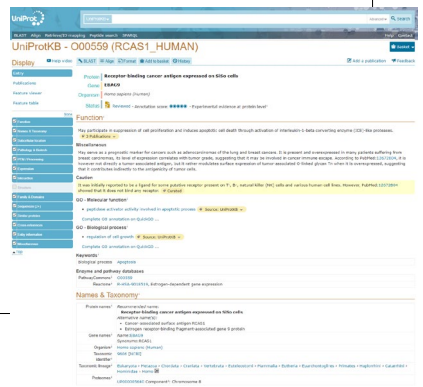
Related documents

Taxonomy

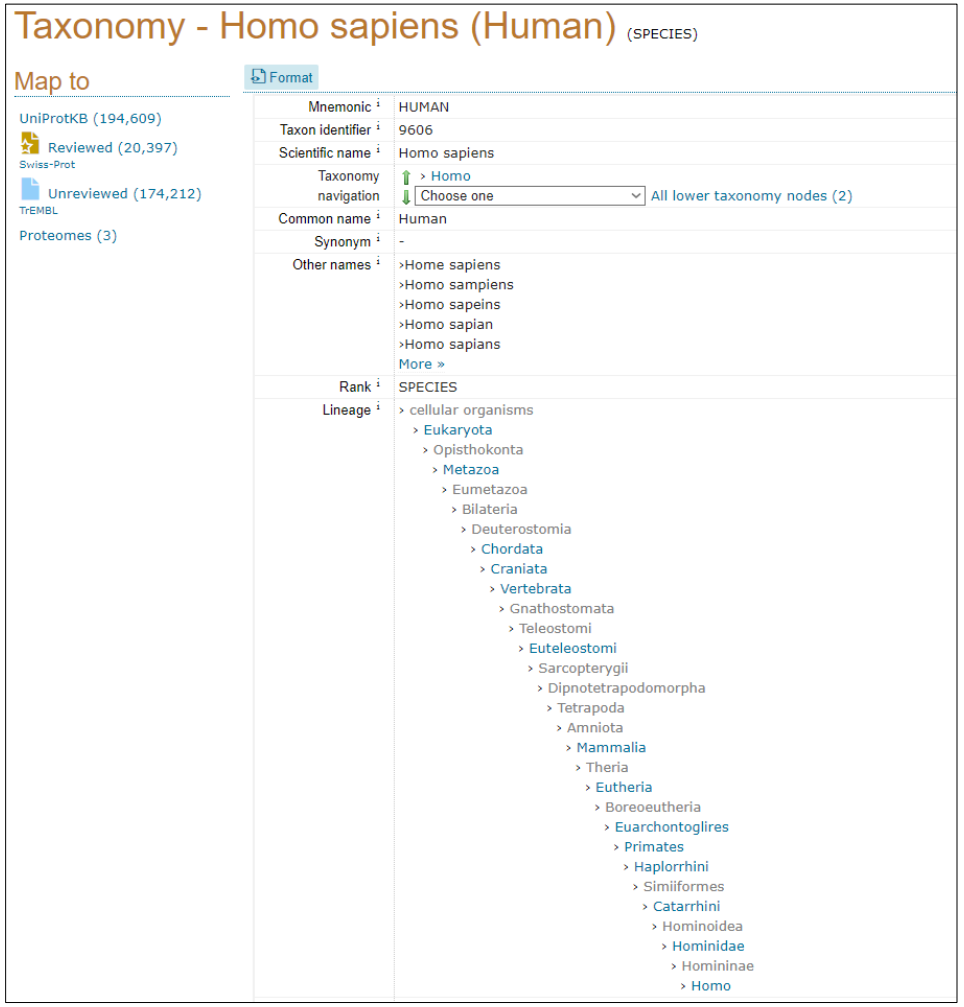
Controlled vocabulary of species

What are proteomes?

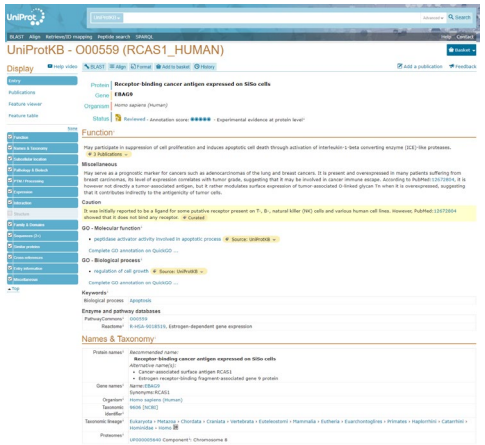
What are reference proteomes?



Vocabularies (I2)



Each metadata **value** comes from a controlled vocabulary – no free form answers.



Qualified References (I3)

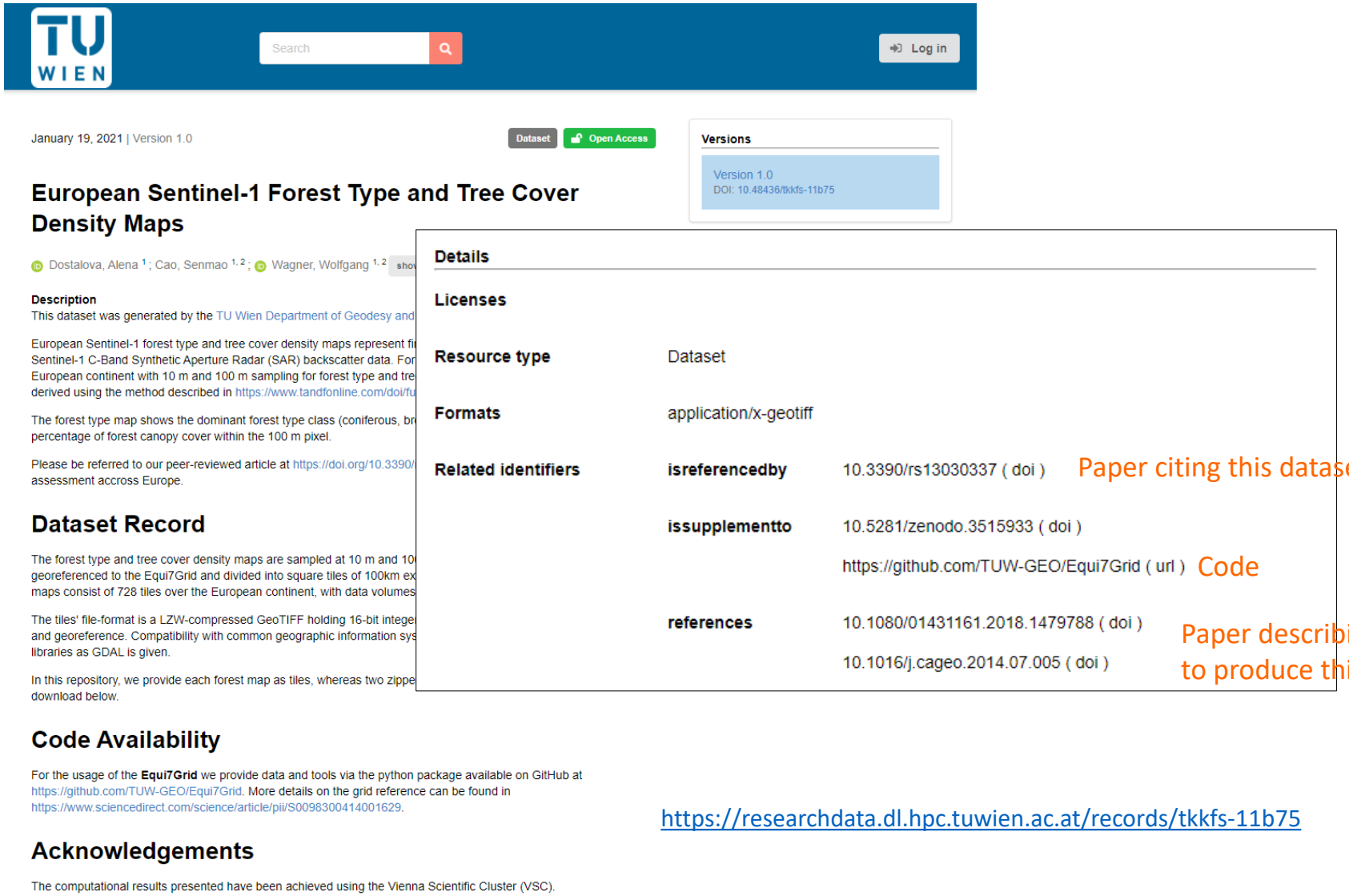
- Meaningul links to describe connections
  - Dataset X was *derived from* dataset Y
  - Dataset Y was *produced* using code Z
- Standard relations define by Data Cite
  - https://schema.datacite.org/meta/kernel-4.3/
- Use persistent identifiers

```
<language>en</language>
<resourceType resourceTypeGeneral="Workflow">Software</resourceType>
<relatedIdentifiers>
  <relatedIdentifier relatedIdentifierType="DOI"
  relationType="IsReferencedBy">10.5072/2047-217X-1-1</relatedIdentifier>
  <relatedIdentifier relatedIdentifierType="DOI"
  relationType="Compiles">10.5072/100038</relatedIdentifier>
</relatedIdentifiers>
<sizes>
  <size>31 MB</size>
</sizes>
```

https://schema.datacite.org/meta/kernel-4.3/example/datacite-example-workflow-v4.xml

IsContinuedBy
Continues
IsDescribedBy
Describes
HasMetadata
IsMetadataFor
HasVersion
IsVersionOf
IsNewVersionOf
IsPreviousVersionOf
IsPartOf
HasPart
IsReferencedBy
References
IsDocumentedBy
Documents
IsCompiledBy
Compiles
IsVariantFormOf
IsOriginalFormOf
IsIdenticalTo
IsReviewedBy
Reviews
IsDerivedFrom
IsSourceOf
IsRequiredBy
Requires
IsObsoletedBy
Obsoletes

Qualified References (I3)



https://researchdata.dl.hpc.tuwien.ac.at/records/tkkfs-11b75

Qualified References (I3)

README.rst

Equi7Grid

build passing | coverage 32% | pypi package 0.0.12 | docs passing

A python class for working with Equi7Grid - how to convert to - how to use the tiling system - etc.

It's a python package that handles the geometric and geographic operations of a gridded and tiled projection system. It was designed for data cubes ingesting satellite imagery and builds the basis for the Equi7Grid (see https://github.com/TUW-GEO/Equi7Grid).

A detailed documentation on the Equi7Grid definition is at:

~/docs/doc_files/

Overlays for visualisation in Google Earth can be found here:

~/docs/doc_files/google_earth_overlays/

Citation

DOI 10.5281/zenodo.1048530

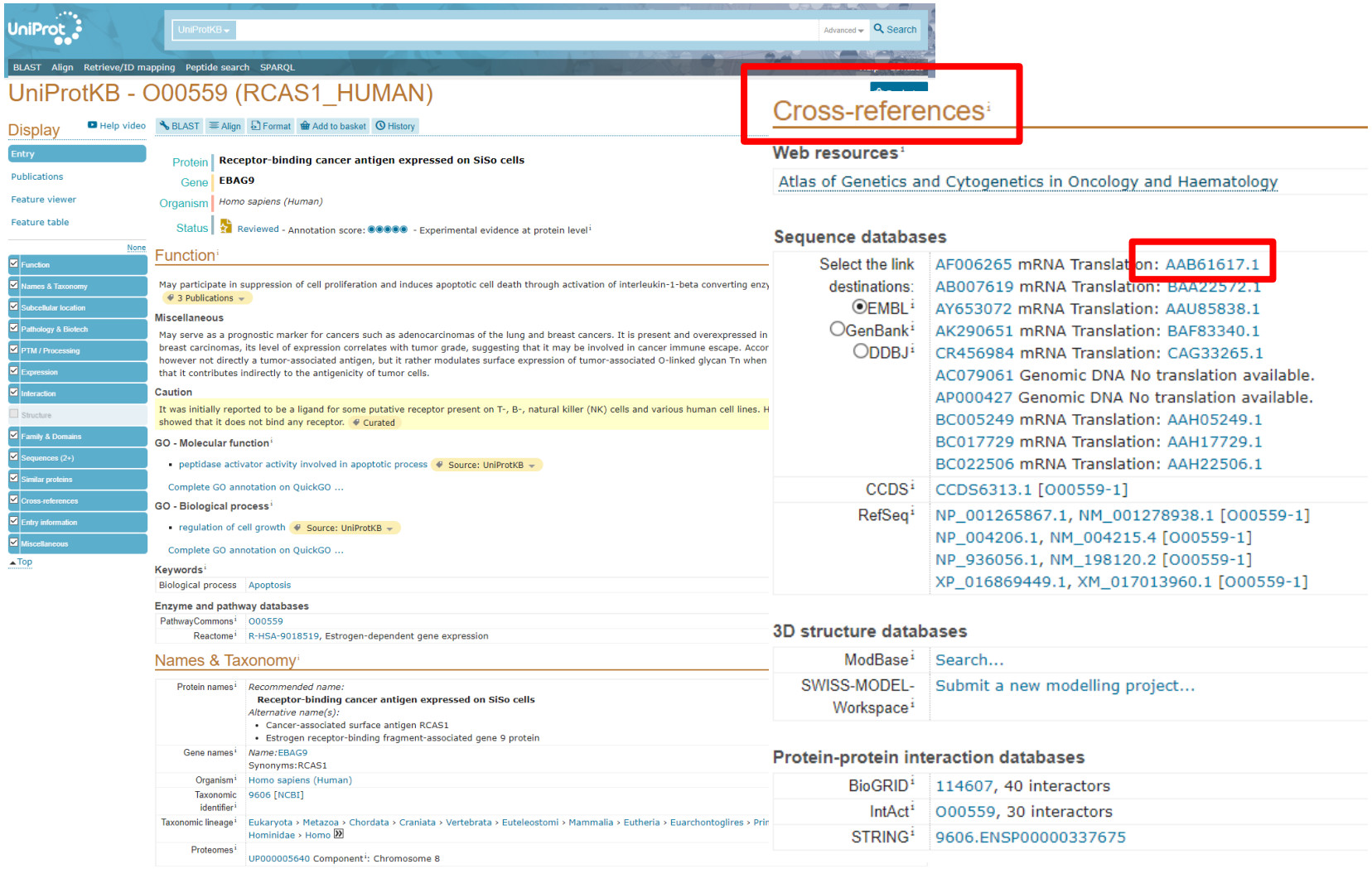
If you use the software in a publication then please cite it using the Zenodo DOI. Be aware that this badge links to the latest package version.

Please select your specific version at https://doi.org/10.5281/zenodo.1048530 to get the DOI of that version. You should normally always use the DOI for the specific version of your record in citations. This is to ensure that other researchers can access the exact research artefact you used for reproducibility.

You can find additional information regarding DOI versioning at http://help.zenodo.org/#versioning

https://github.com/TUW-GEO/Equi7Grid

Qualified References (I3)



Reusable

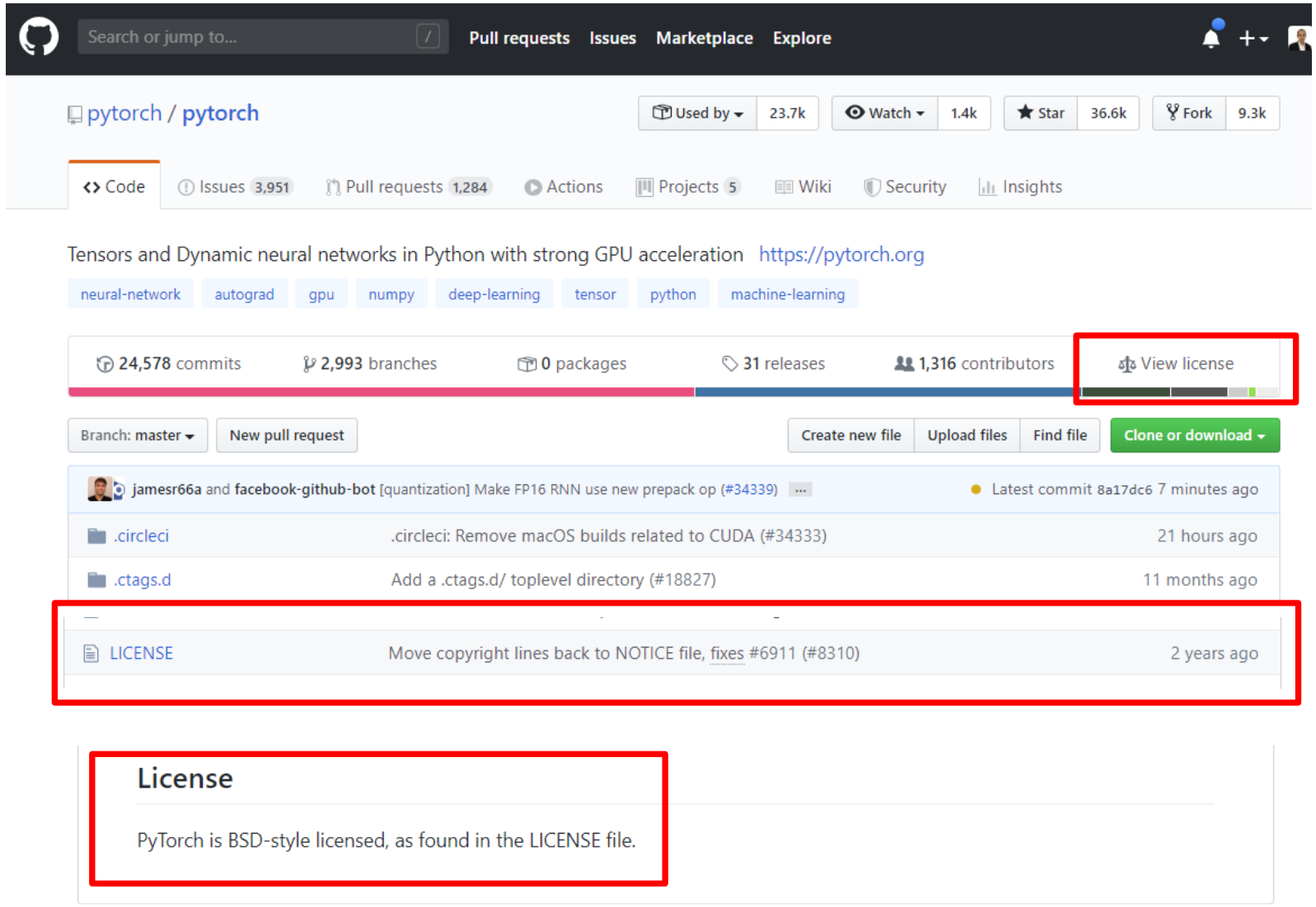
- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage **license**
  - R1.2. (Meta)data are associated with detailed **provenance**
  - R1.3. (Meta)data meet domain-relevant **community standards**

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

FAIR OFFICE AUSTRIA

- Public repository on GitHub
  - May suggest that authors are willing to share code
- No license
  - no possibility for reuse
  - can only by viewed (only because terms of use enforce that)
- Code without a license is like an object in a museum
  - You can watch and admire it, but you cannot touch it!

License (R1.1)


Search or jump to...
Pull requests
Issues
Marketplace
Explore
pytorch / pytorch
Used by 23.7k
Watch 1.4k
Star 36.6k
Fork 9.3k
Code
Issues 3,951
Pull requests 1,284
Actions
Projects 5
Wiki
Security
Insights
Tensors and Dynamic neural networks in Python with strong GPU acceleration https://pytorch.org
neural-network
autograd
gpu
numpy
deep-learning
tensor
python
machine-learning
24,578 commits
2,993 branches
0 packages
31 releases
1,316 contributors
View license
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
jamesr66a and facebook-github-bot [quantization] Make FP16 RNN use new prepack op (#34339)
Latest commit 8a17dc6 7 minutes ago
.circleci
.circleci: Remove macOS builds related to CUDA (#34333)
21 hours ago
.ctags.d
Add a .ctags.d/ toplevel directory (#18827)
11 months ago
LICENSE
Move copyright lines back to NOTICE file, fixes #6911 (#8310)
2 years ago
License
PyTorch is BSD-style licensed, as found in the LICENSE file.

R1.2 (Meta)data are associated with detailed provenance

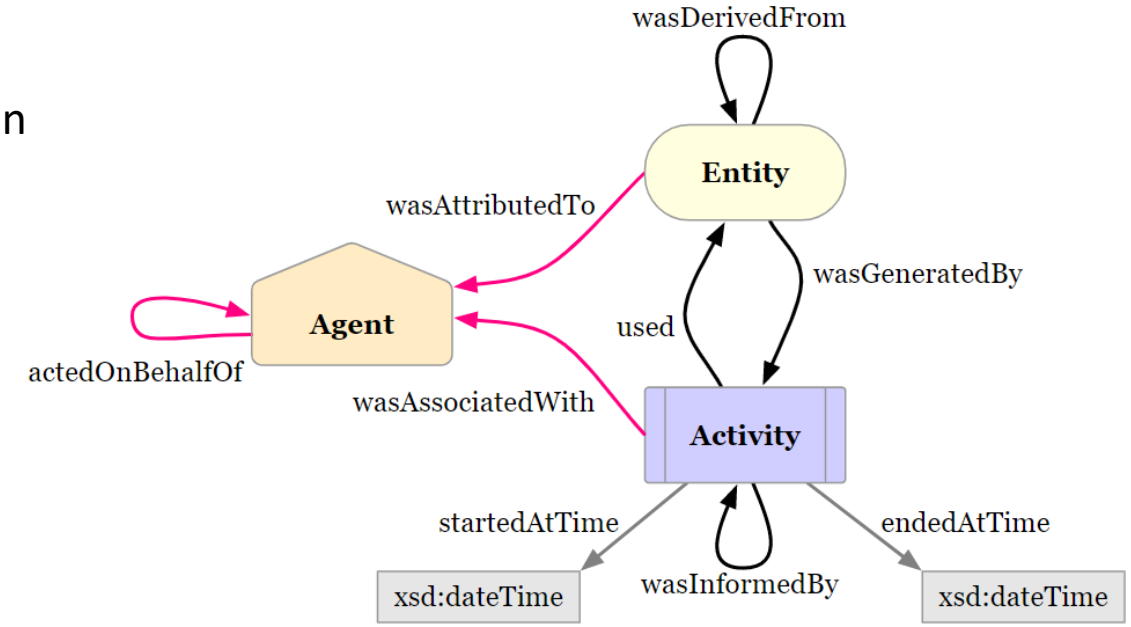
FAIR OFFICE AUSTRIA

- Provenance
  - Describes origin of data
  - Who? What? When? How?
- Supports evaluation and can build trust in data
  - 'Officially, North Korea claims to have identified zero cases of COVID-19 inside its territory' https://www.npr.org/sections/goatsandsoda/2020/02/20/807027901/north-korea-claims-zero-cases-of-coronavirus-infection-but-experts-are-skeptical?t=1615196582563

R1.2 (Meta)data are associated with detailed provenance

FAIR OFFICE AUSTRIA

- PROV-O: The PROV Ontology
  - Machine-actionable way to express provenance
- Licenses (level of compliance for repository preservation /policy/strategy/action plan)
- To what extend the data is useful for others.
- Rich Metadata that describe the context under which the data was generated.
- What usage rights do you attach to your data?
- Provenance means to describe origin of data (Who? What? When? How?) → Ontology



https://www.w3.org/TR/prov-o/

R1.3. (Meta)data meet domain-relevant community standards

FAIR OFFICE AUSTRIA

- Who is the “community” ?
- What is the “standard” ?
  - English vs other languages
- Metadata
  - Domain independent
    - e.g. Dublin Core
  - Domain specific
    - e.g. EXIF for images
- Sometimes no common standard exist
  - Good documentation and README
- There is no universal guideline – it always depends!

R1.3. (Meta)data meet domain-relevant community standards

FAIR OFFICE AUSTRIA

- Does not have to be a standard for everyone!

COVID-19 Gs

3G-Regel: 3G steht für geimpft, genesen oder getestet. Beim Test gelten SOWOHL Antigentests ALS AUCH PCR-Tests. Sind also 3G für einen Zutritt nötig, muss man ENTWEDER geimpft, genesen ODER getestet sein.

2G-Regel: 2G steht für geimpft oder genesen. Werden für einen Zutritt 2G gefordert, dann muss man ENTWEDER geimpft ODER genesen sein.

2G+ oder 2Gplus steht für geimpft oder genesen PLUS getestet, und zwar mit einem PCR-Test. Wird für einen Zutritt die 2G+-Regel verlangt, muss man also ENTWEDER geimpft ODER genesen UND ZUSÄTZLICH getestet sein, und zwar mit einem PCR-Test.

2,5G-Regel: 2,5G steht für geimpft oder genesen oder PCR-getestet. Wird für einen Zutritt also 2,5G verlangt, dann muss man ENTWEDER geimpft, genesen ODER PCR-getestet sein.

NATO Phonetic Alphabet

A	Alpha	N	November
B	Bravo	O	Oscar
C	Charlie	P	Papa
D	Delta	Q	Quebec
E	Echo	R	Romeo
F	Foxtrot	S	Sierra
G	Golf	T	Tango
H	Hotel	U	Uniform
I	India	V	Victor
J	Juliet	W	Whiskey
K	Kilo	X	X-ray
L	Lima	Y	Yankee
M	Mike	Z	Zulu

R1.3. (Meta)data meet domain-relevant community standards

FAIR OFFICE AUSTRIA

- Good documentation supports reuse
  - Removes ambiguities (especially where there are no common controlled vocabularies or others standards)
- Example
  - Confirmed cases of COVID-19: testing date vs reporting date

Indicators	Definition
Tests	Cumulative number of tests carried out for SARS-CoV-2, from 27 February 2020 up to and including the reporting date. Responsible for data consolidation: Office of the respective federal state government (Land), data status: morning of the reporting day.
Laboratory-confirmed cases	Cumulative number of laboratory-confirmed cases of SARS-CoV-2 infection (sum of "Active cases", "Recovered cases" and "Deceased cases") with laboratory diagnosis date since 27.02.2020 up to and including the reporting date.
Active cases	Cumulative number of laboratory-confirmed cases of SARS-CoV-2 infection with laboratory diagnosis date from 27.02.2020 up to and including the reporting date, which have not been classified as "recovered" or "deceased" on the reporting date.
Recovered cases	Cumulative number of laboratory-confirmed cases of SARS-CoV-2 infection with laboratory diagnosis date from 27.02.2020 up to and including the report date, which are classified as "recovered" on the report date. Definition of "recovered" (since 9 July): in the case of home care, 10-day home isolation after the onset of symptoms or laboratory diagnosis; in case of severe disease progression, the earliest 10 days after onset of symptoms, at least 48 hours without symptoms AND the following result by RT-PCR according to the Charité protocol: no nucleic acid detection of beta-coronavirus SARS-CoV-2 or nucleic acid detection of beta-coronavirus SARS-CoV-2 at a Ct value of more than 30. Further details can be found in the recommendation for the release of COVID-19 cases, recommendation for the release of COVID-19 cases from isolation.
Deceased cases	Cumulative number of laboratory-confirmed cases of SARS-CoV-2 infection with a laboratory diagnosis date from 27.02.2020 up to and including the report date, which are classified as "deceased" on the report date.

https://covid19-dashboard.ages.at/basisinfo_en.html?l=en

To summarize FAIR Principles…

- Data can be **Findable** by adding metadata and a persistent identifier.
- Data can be **Accessible** by defining who can access data and how, but keep in mind that, if you can't publish data openly, you should provide access to the metadata, for example through a data repository.
- Data can be **Interoperable by** using common standards and open data formats.
- Data can be **Reusable** by adding documentation that help others understand data and an appropriate data license that determines how data can be reused.

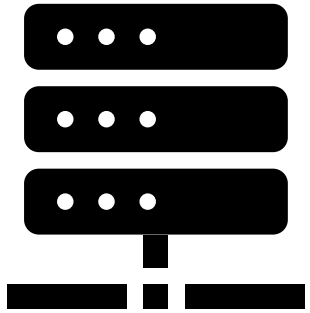
How repositories support FAIRness

Repositories and FAIRness

Data repositories are key in putting the FAIR principles into practice.

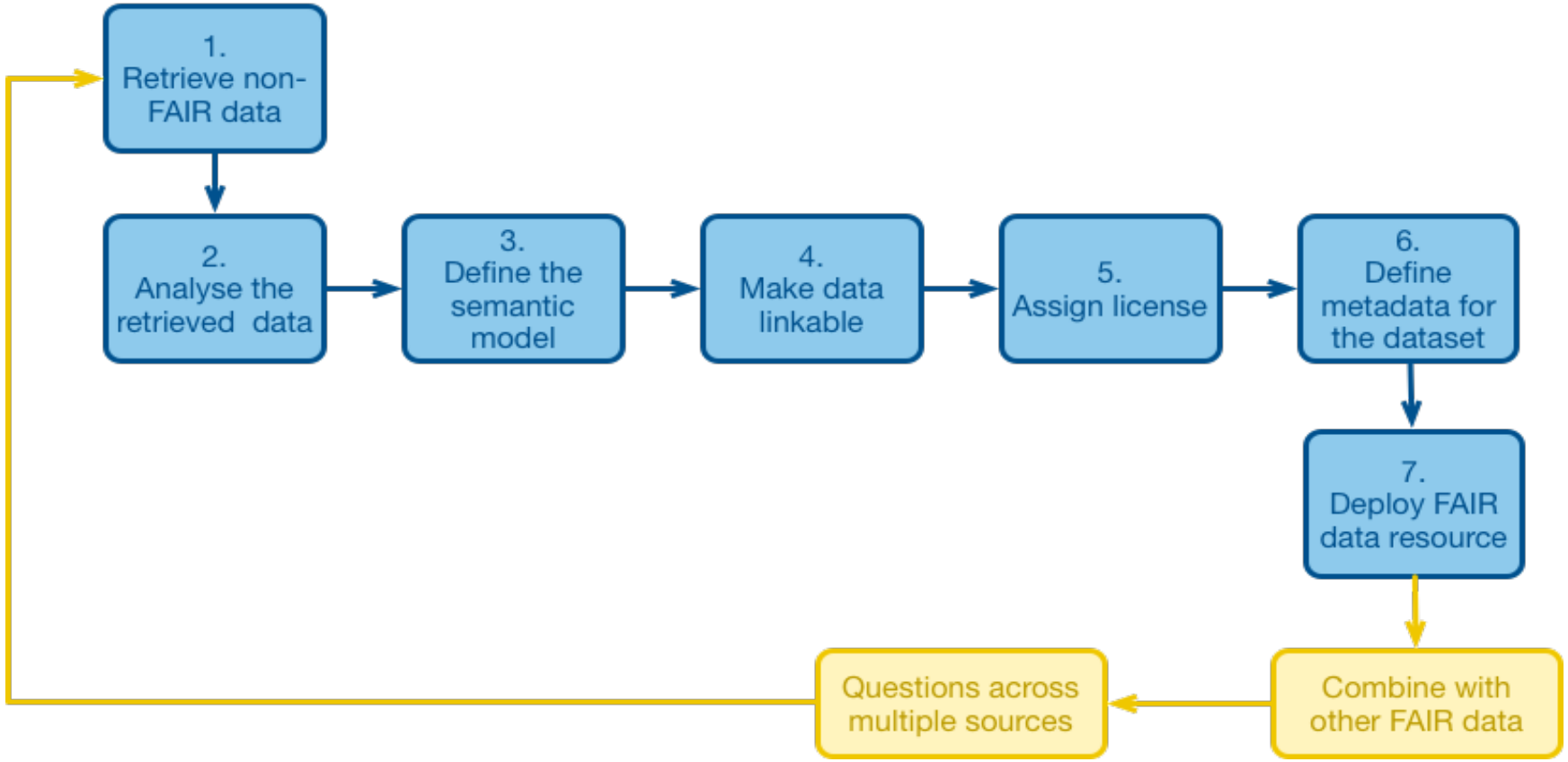
Not only do they enable findability and access, but they also provide persistent identifiers, documentation and metadata, thus fostering reusability for humans and machines.

- A “form” needs to be filled –metadata by default.
- The form complies with a specific metadata standard.
- Metadata will then become machine-actionable and searchable in an online resource.
- A persistent identifier for the data is automatically generated.
- References to other data or metadata can be included.
- Authentication and authorization procedures are in place.
- Access can be regulated from closed to open.
- The provision of machine-readable licenses enhances the reusability of the data.
- The use of standards and controlled vocabularies is enforced.
- Interfaces for external services like OAI-PMH allow harvesting of metadata for stored records.



How to FAIRify data?

FAIRification Process

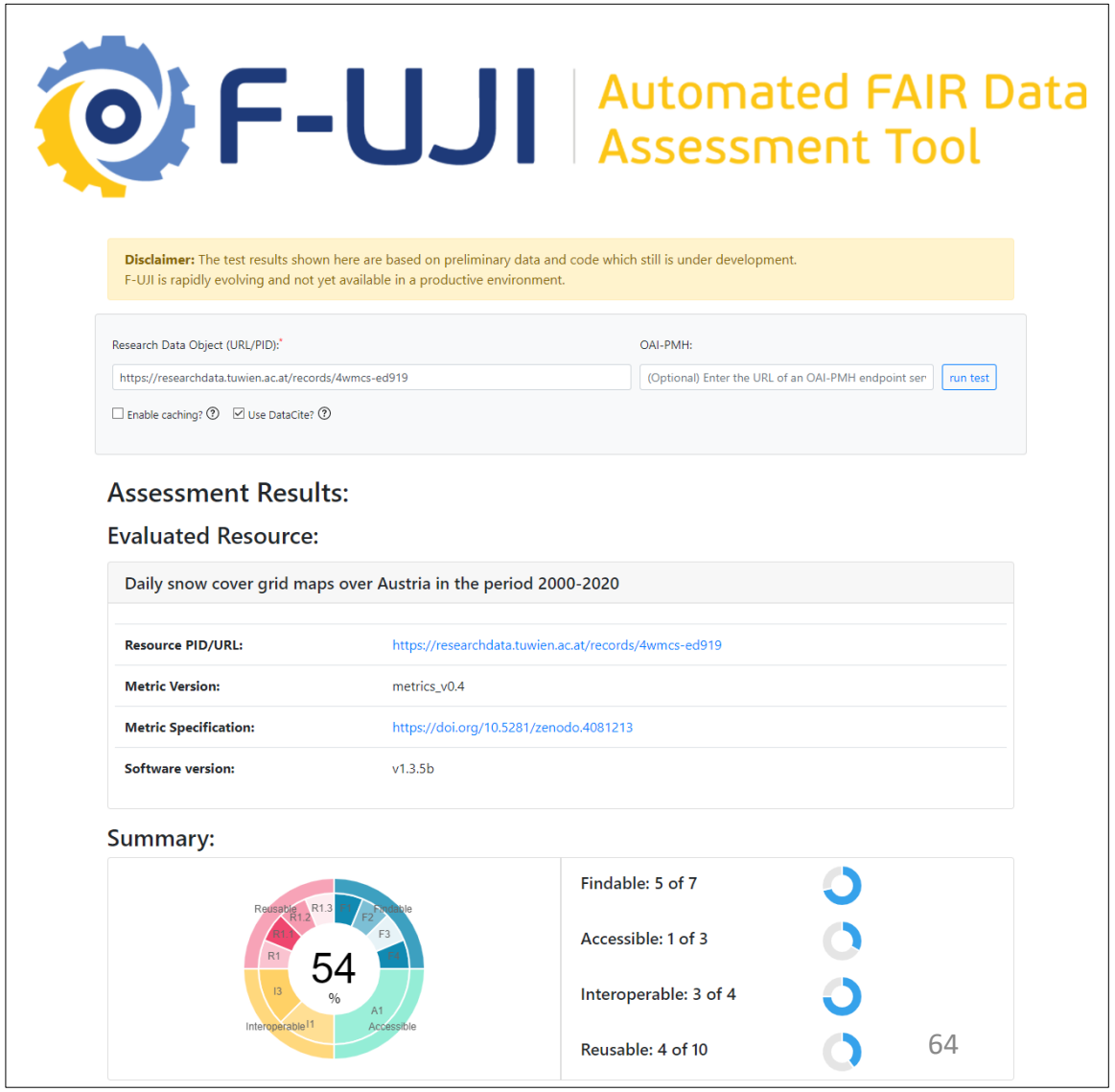


FAIRification Process - GO FAIR (go-fair.org)

FAIR Assessment - F-UJI Tool

- F-UJI is a web service to programmatically assess FAIRness of research data objects based on metrics developed by the FAIRsFAIRproject.

https://f-uji.net



Report:

Findable

- FsF-F1-01D - Data is assigned a globally unique identifier.
- FsF-F1-02D - Data is assigned a persistent identifier.
- FsF-F2-01M - Metadata includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability.
- FsF-F3-01M - Metadata includes the identifier of the data it describes.
- FsF-F4-01M - Metadata is offered in such a way that it can be retrieved programmatically.

Accessible

- FsF-A1-01M - Metadata contains access level and access conditions of the data.
- FsF-A1-03D - Data is accessible through a standardized communication protocol.
- FsF-A1-02M - Metadata is accessible through a standardized communication protocol.

Interoperable

- FsF-I1-01M - Metadata is represented using a formal knowledge representation language.
- FsF-I1-02M - Metadata uses semantic resources
- FsF-I3-01M - Metadata includes links between the data and its related entities.

Reusable

- FsF-R1-01MD - Metadata specifies the content of the data.
- FsF-R1.1-01M - Metadata includes license information under which data can be reused.
- FsF-R1.2-01M - Metadata includes provenance information about data creation or generation.
- FsF-R1.3-01M - Metadata follows a standard recommended by the target research community of the data.
- FsF-R1.3-02D - Data is available in a file format recommended by the target research community.

FAIR OFFICE AUSTRIA

Metric tests:

Test:	Test name:	Result:
FsF-R1-01MD-1	Minimal information about available data content is given in metadata	✓
FsF-R1-01MD-1a	Resource type (e.g. dataset) is given in metadata	✓
FsF-R1-01MD-1b	Information about data content (e.g. links) is given in metadata	✗
FsF-R1-01MD-2	Verifiable data descriptors (file info, measured variables or observation types) are specified in metadata	✗
FsF-R1-01MD-2a	File size and type information are specified in metadata	✗
FsF-R1-01MD-2b	Measured variables or observation types are specified in metadata	✗
FsF-R1-01MD-3	Data content matches file type and size specified in metadata	✗
FsF-R1-01MD-4	Data content matches measured variables or observation types specified in metadata	✗

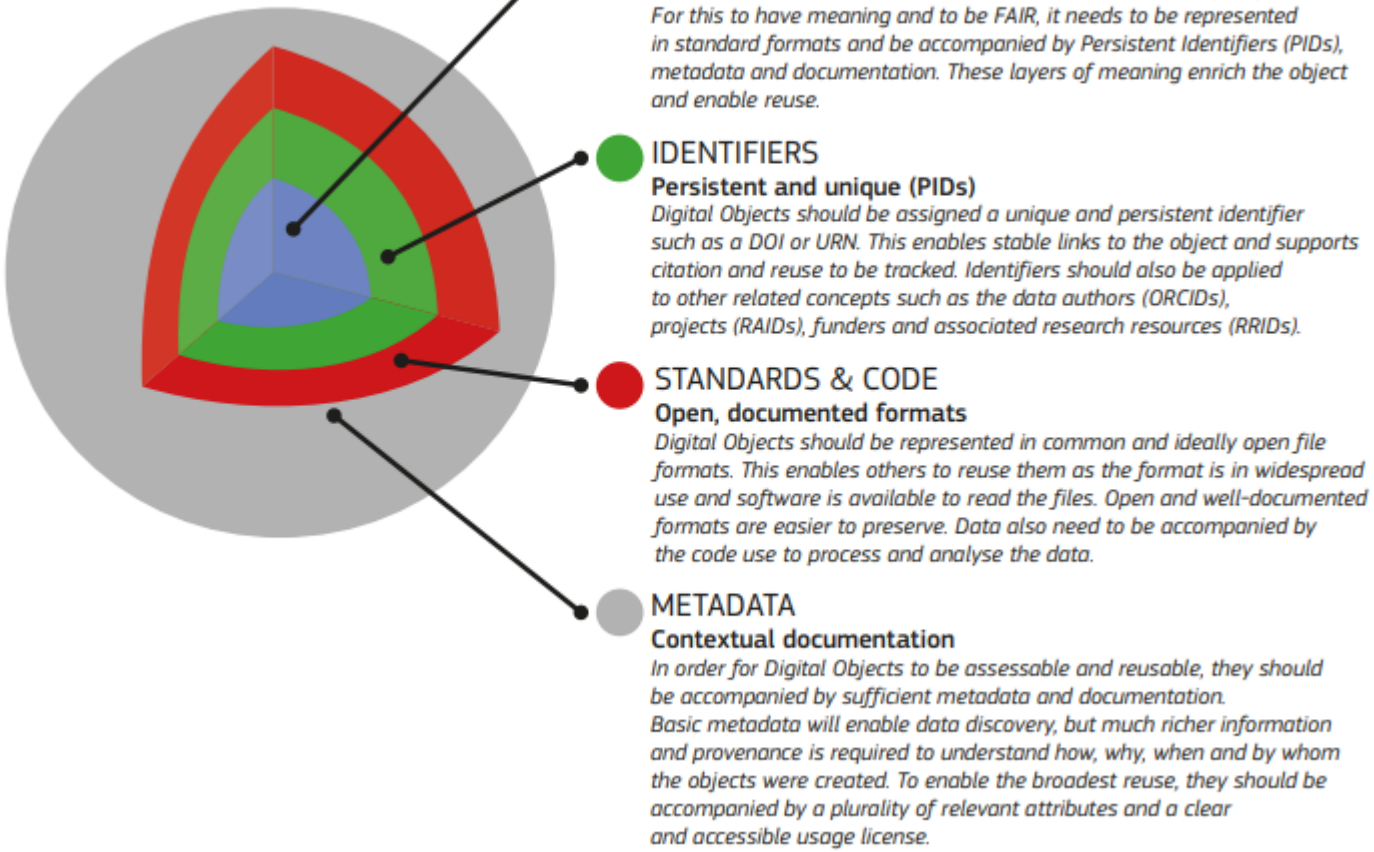
Debug:

Level:	Message:
INFO	Object landing page accessible status -: True
SUCCESS	Resource type specified -: dataset
WARNING	NO data object content available/accessible to perform file descriptors (type and size) tests
WARNING	NO measured variables found in metadata, skip 'measured_variable' test.
WARNING	Measured variables given in metadata do not match data object content

Code still under development!

https://www.f-uji.net/index.php

FAIR Digital Object (FAIR ecosystem)



Turning FAIR into reality https://op.europa.eu/s/oM5N

GO FAIR

GO FAIR?

- Community that provides Practical Guidance to support research stockholders to implement the FAIR data Principles.

GO FAIR Framework

- Three-point FAIRification framework provides practical guidnance to stackholders seeking to go FAIR.

- Metadata for Machines workshops (M4M)
- FAIR Implementation Profiles (FIP)
- FAIR Data Point (FDP)

GO FAIR Implementation Network

- Work groups to establish specific materials and tools as elements of the Internet of FAIR data and Services (IFDS).

GO FAIR : How to join?

- You can do so either by joining an **existing Implementation Network** or by forming a new one.
- The INs are active in three activity pillars: **GO CHANGE**, **GO TRAIN** and **GO BUILD**
- IN Example: **Food Systems**

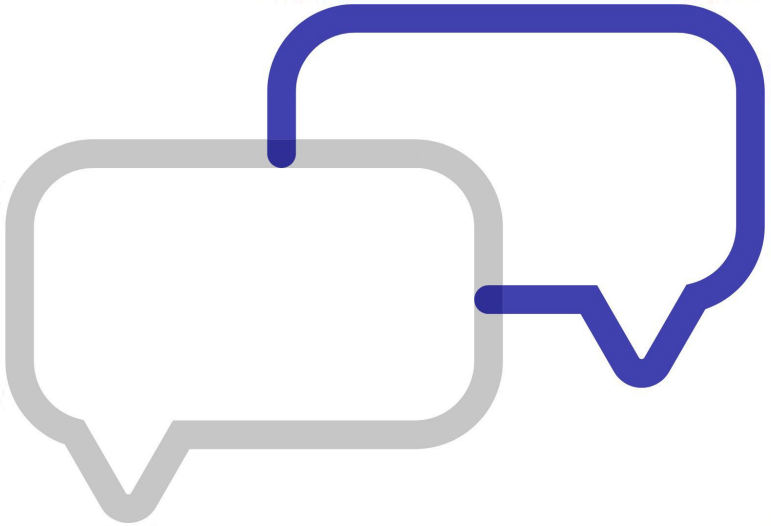
https://www.go-fair.org/

FAIR Office Austria

FAIR Office Austria

- Founding team = Consortium TU Wien, Graz University of Technology & University of Vienna
- Mission: **We connect stakeholders from research communities and service providers. Together, we help to advance the FAIR principles.**
- Support of researchers at partner institutions
  - FAIR in project proposals and DMPs
  - FAIR in research processes
  - FAIR through optimal use of tools and services
  - Support from data stewards
- Support of service providers
  - FAIRisation of repositories
  - Enhancing automation and machine actionability
- Partner with EOSC Mandated Organization and RDA Austria
  - Exchange with national and international initiatives
- Monitoring of FAIR processes in Austria
- Since June 2021: FAIR Office Austria is GO FAIR National Support and Coordination Office
- More information see: https://fair-office.at/?lang=en

Q&A



FAIR
OFFICE AUSTRIA

Contact

contact@fair-office.at
samah.jaber@tuwien.ac.at

This work is licensed under the Creative Commons Attribution 4.0 International License unless otherwise stated.

FAIR
OFFICE AUSTRIA

Thank you!