Open research data for use and re-use

Ana Alice Baptista
https://www.cienciavitae.pt/201D-B2FC-E126

Braga, July / 2023

Universidade do Minho
Escola de Engenharia

Contents

- Open Science
- Open Data
- FAIR Principles
- Re-use cases

Open Science

“An approach to the scientific process that focuses on spreading knowledge as soon as it is available using digital and collaborative technology”

“Open Science.” *Research-And-Innovation.ec.europa.eu*, research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en.

"Knowledge is of all and for all"

“SOBRE CIÊNCIA ABERTA.” *Ciencia Aberta*, Ministério da Ciência, Tecnologia e Ensino Superior, 2016, www.ciencia-aberta.pt/sobre-ciencia-aberta. Accessed 16 June 2023.

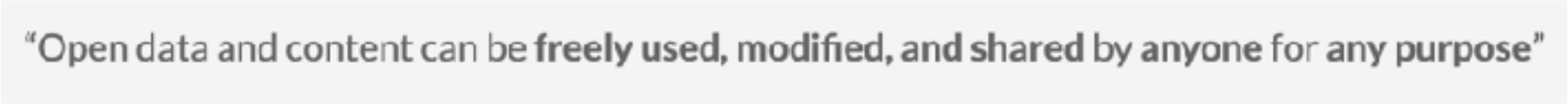
Ambitions of the EU's Open Science Policy

- Open Data
- European Open Science Cloud
- New generation metrics
- Future of scholarly communication
- Rewards
- Integrity & Reproducibility
- Education & Skills
- Citizen Science

"Open Science." *Research-And-Innovation.ec.europa.eu*, research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en.

Open Definition

Open Knowledge
FOR A FAIR, SUSTAINABLE AND OPEN FUTURE

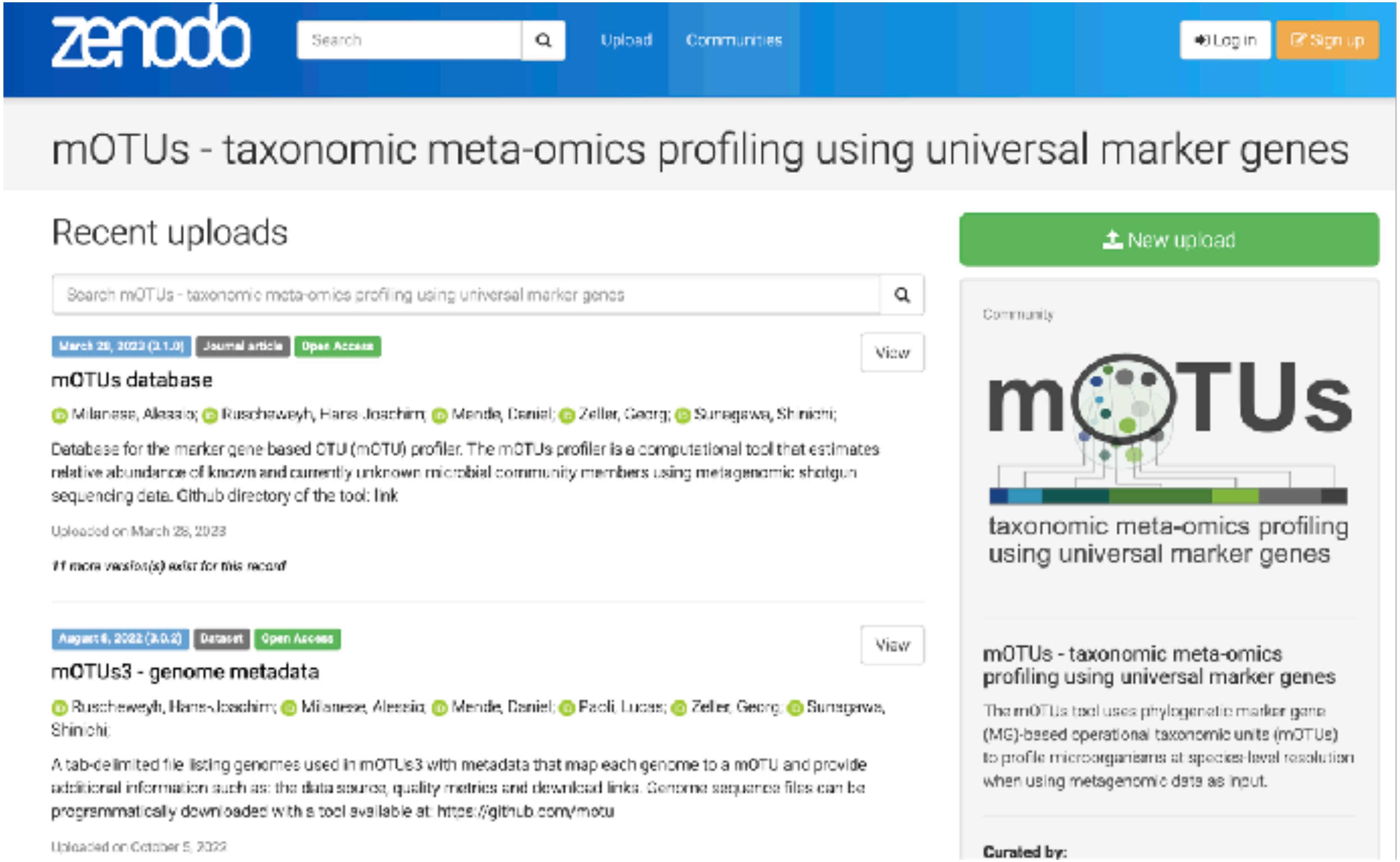


"Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**"

Open Knowledge Foundation. "The Open Definition - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge." *Opendefinition.org*, opendefinition.org/.

"Open research data refers to the data underpinning scientific research results that has no restrictions on its access, enabling anyone to access it"

"Facts and Figures for Open Research Data." *Research-And-Innovation.ec.europa.eu*, Directorate-General for Research and Innovation, research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/open-science-monitor/facts-and-figures-open-research-data_en. Accessed 16 June 2023.


zenodo
Search
Upload
Communities
Log in
Sign up

mOTUs - taxonomic meta-omics profiling using universal marker genes

Recent uploads

Search mOTUs - taxonomic meta-omics profiling using universal marker genes

March 28, 2023 (3.1.0) Journal article Open Access

mOTUs database

Milanese, Alessio; Ruscheweyh, Hans Joachim; Mende, Daniel; Zeller, Georg; Sunagawa, Shinichi;

Database for the marker gene based OTU (mOTU) profiler. The mOTUs profiler is a computational tool that estimates relative abundance of known and currently unknown microbial community members using metagenomic shotgun sequencing data. Github directory of the tool: link

Uploaded on March 28, 2023

11 more version(s) exist for this record

August 6, 2022 (3.0.2) Dataset Open Access

mOTUs3 - genome metadata

Ruscheweyh, Hans-Joachim; Milanese, Alessio; Mende, Daniel; Paoli, Lucas; Zeller, Georg; Sunagawa, Shinichi;

A tab-delimited file listing genomes used in mOTUs3 with metadata that map each genome to a mOTU and provide additional information such as: the data source, quality metrics and download links. Genome sequence files can be programmatically downloaded with a tool available at: https://github.com/motu

Uploaded on October 5, 2022

mOTUs - taxonomic meta-omics profiling using universal marker genes

The mOTUs tool uses phylogenetic marker gene (MG)-based operational taxonomic units (mOTUs) to profile microorganisms at species-level resolution when using metagenomic data as input.

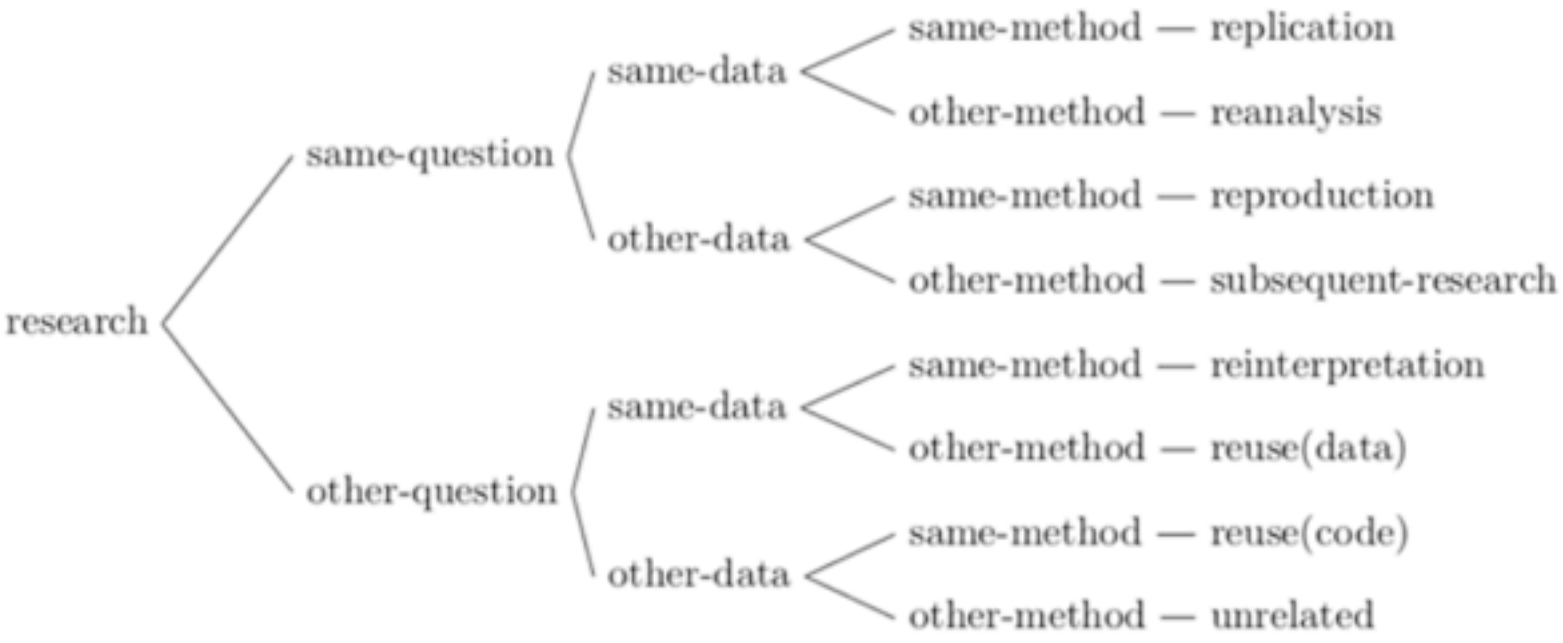
Curated by:

"MOTUs - Taxonomic Meta-Omics Profiling Using Universal Marker Genes." *Zenodo*, zenodo.org/communities/motus/. Accessed 16 June 2023.

Why open research data?

- Because it is compulsory (mandates)
- For transparency & reproducibility
- For public scrutiny
- For trustability
- For use & re-use

What may open research data be used for?


research
same-question
other-question
same-data
other-data
same-data
other-data
same-method — replication
other-method — reanalysis
same-method — reproduction
other-method — subsequent-research
same-method — reinterpretation
other-method — reuse(data)
same-method — reuse(code)
other-method — unrelated

Van de Sandt, S., Dallmeier-Tiessen, S., Lavasa, A. and Petras, V., 2019. The Definition of Reuse. Data Science Journal, 18(1), p.22. DOI: http://doi.org/10.5334/dsj-2019-022

Why not use Open Research Data?

- Operational:
  - Language
  - Format
  - up-to-dateness
  - Meaning (variable, units of measure,…)
  - License
- Cultural & Political:
  - Replication/reinterpretation/reanalysis studies not so sexy
  - Trust / reliability
  - Culture
  - No explicit rewards

For the data to be used & re-used it has to be made available the right way

FAIR principles

Findable

Accessible

Interoperable

Re-usable

FAIR

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

FAIR

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

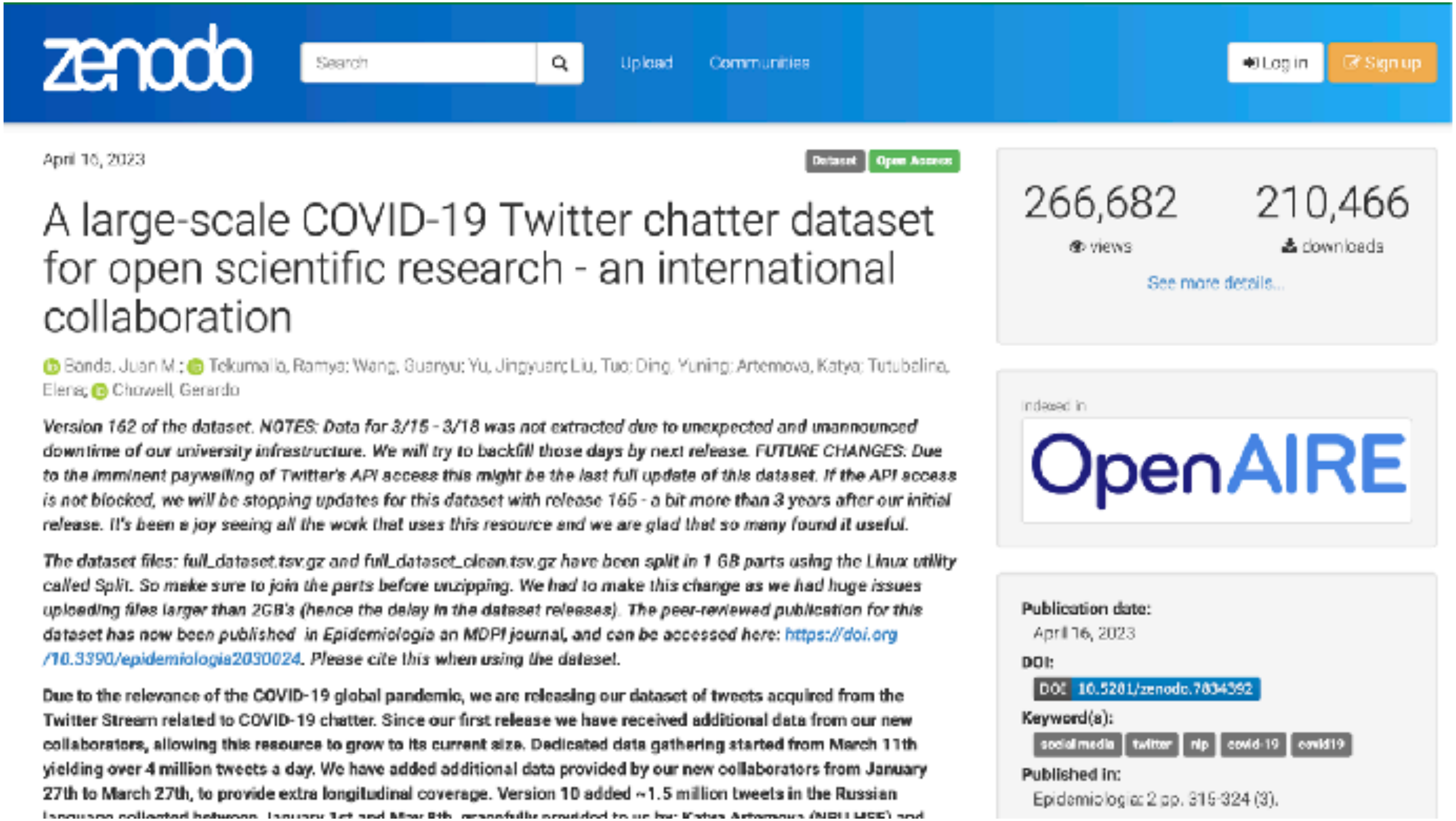
R1.1: (meta)data are released with a clear and accessible data usage license

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

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

https://about.zenodo.org/principles/

REUSE cases



April 16, 2023

A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration

Banda, Juan M.; Tekumalla, Ramya; Wang, Guanyu; Yu, Jingyuan; Liu, Tuo; Ding, Yuning; Artemova, Katya; Tutubalina, Elena; Chowell, Gerardo

Version 162 of the dataset. NOTES: Data for 3/15 - 3/18 was not extracted due to unexpected and unannounced downtime of our university infrastructure. We will try to backfill those days by next release. FUTURE CHANGES: Due to the imminent paywalling of Twitter's API access this might be the last full update of this dataset. If the API access is not blocked, we will be stopping updates for this dataset with release 165 - a bit more than 3 years after our initial release. It's been a joy seeing all the work that uses this resource and we are glad that so many found it useful.

The dataset files: full_dataset.tsv.gz and full_dataset_clean.tsv.gz have been split in 1 GB parts using the Linux utility called Split. So make sure to join the parts before unzipping. We had to make this change as we had huge issues uploading files larger than 2GB's (hence the delay in the dataset releases). The peer-reviewed publication for this dataset has now been published in Epidemiologia an MDPI journal, and can be accessed here: https://doi.org/10.3390/epidemiologia2030024. Please cite this when using the dataset.

Due to the relevance of the COVID-19 global pandemic, we are releasing our dataset of tweets acquired from the Twitter Stream related to COVID-19 chatter. Since our first release we have received additional data from our new collaborators, allowing this resource to grow to its current size. Dedicated data gathering started from March 11th yielding over 4 million tweets a day. We have added additional data provided by our new collaborators from January 27th to March 27th, to provide extra longitudinal coverage. Version 10 added ~1.5 million tweets in the Russian language collected between January 1st and May 8th, gracefully provided to us by: Katya Artemova (NRU HSE) and

266,682
views

210,466
downloads

Publication date:
April 16, 2023

DOI:
DOI 10.5281/zenodo.7834392

Keyword(s):
social media | twitter | nlp | covid-19 | covid19

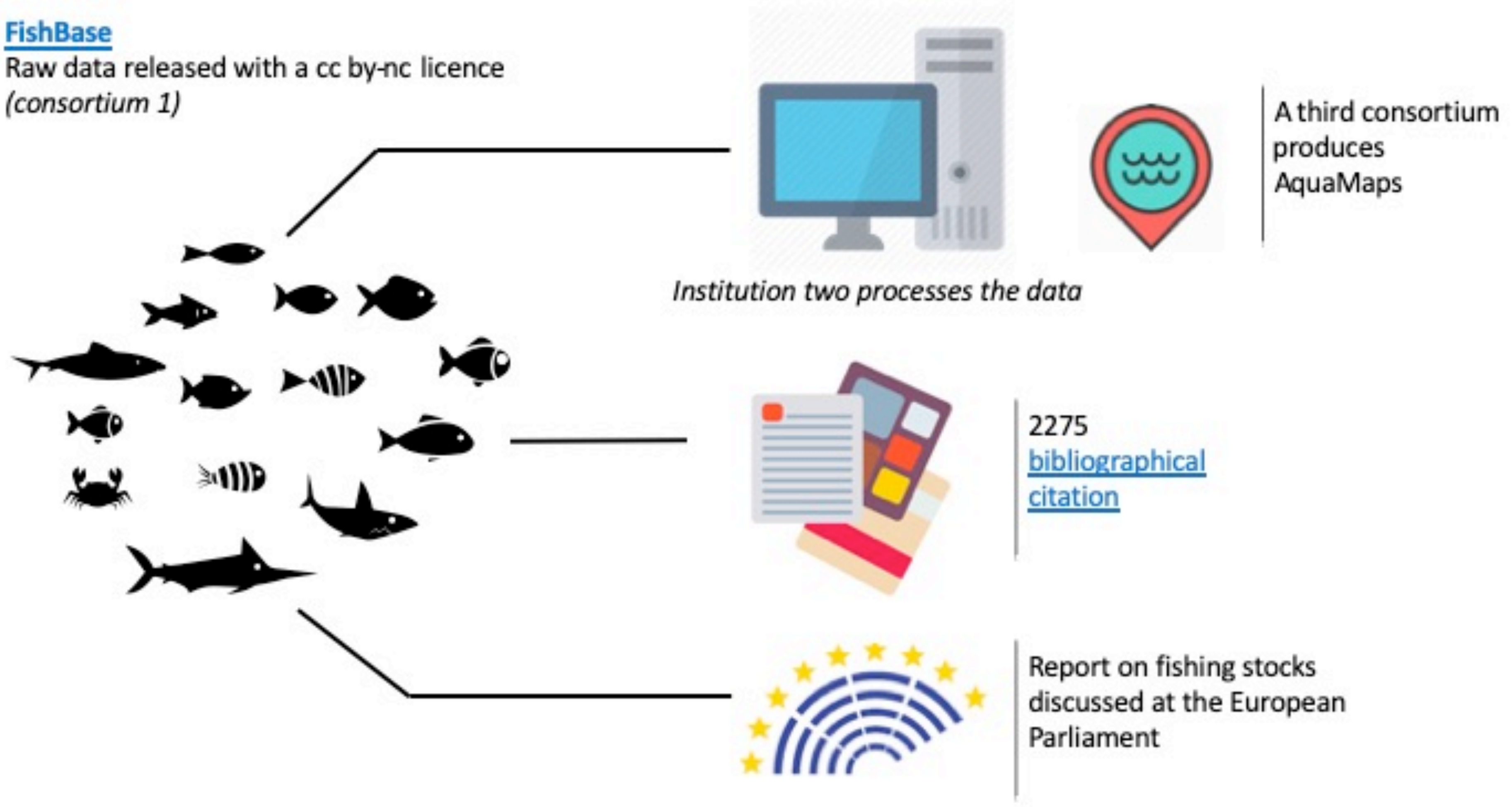
Published in:
Epidemiologia: 2 pp. 315-324 (3).

Banda, Juan M., et al. “A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research - an International Collaboration.” *Zenodo*, 29 Jan. 2023, zenodo.org/record/7582648. Accessed 16 June 2023.

Citizen science in the surveillance and monitoring of mosquito-borne diseases

- Cost reduction for mosquito monitoring systems
- Early warning system for mosquito-borne diseases
- Detections further from the known invasion area

European Commission, Directorate-General for Research and Innovation, Switters, J., Osimo, D., *Citizen science in the surveillance and monitoring of mosquito-borne diseases – Open science monitor case study*, Publications Office, 2019, https://data.europa.eu/doi/10.2777/431775



FishBase
Raw data released with a cc by-nc licence
(consortium 1)

Institution two processes the data

A third consortium
produces
AquaMaps

2275
bibliographical
citation

Report on fishing stocks
discussed at the European
Parliament

“ FishBase is a digital catalogue of fishes, collecting a variety of data on 34,300 fish species, such as geographical distribution, biometrics, habitats, population dynamics as well as reproductive, metabolic and genetic data.”

Pavone, Gina. “Data Reuse Stories. Some Concrete Cases Involving Several Institutions and Consortia in Europe.” *OpenAIRE*, 30 Nov. 2020, www.openaire.eu/blogs/data-reuse-stories-some-concrete-cases-involving-several-institutions-and-consortia-in-europe. Accessed 16 June 2023.

to know more

EUROPEAN OPEN SCIENCE CLOUD

Your unified access to the European hub of research data, tools and services for innovation and education

zenodo

FAIR Data Maturity Model

Specification and Guidelines

2020

data.europa.eu - The official portal for European data

OpenAIRE

Data Catalog Vocabulary (DCAT)

Open science

Making science more accessible, inclusive and equitable for the benefit of all

The Use Case Observatory

A 3-year monitoring of 30 reuse cases to understand the economic, governmental, social and environmental impact of open data

Data on the Web Best Practices

W3C Recommendation 31 January 2017

W3C

Open research data for use and re-use

Ana Alice Baptista
https://www.cienciavitae.pt/201D-B2FC-E126

Braga, July / 2023

Universidade do Minho
Escola de Engenharia