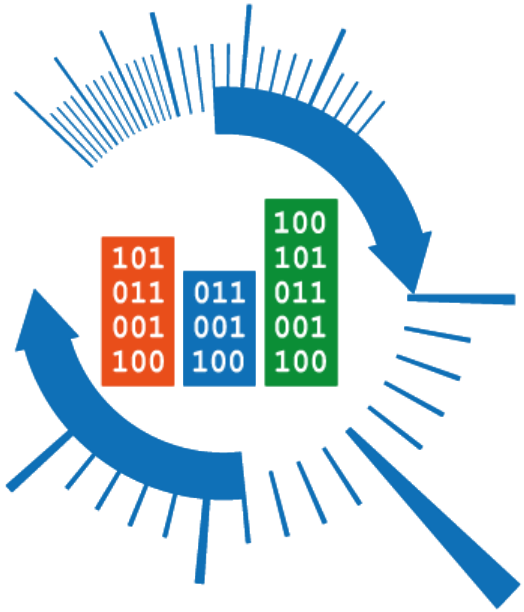
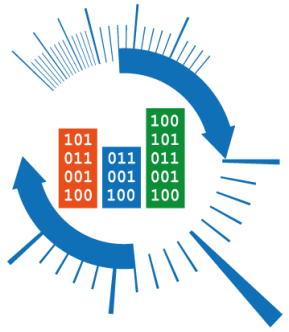


Web Tutorial 2: Planning and Data Publication

This work is licensed under a Creative Commons Attribution 4.0 International (CC-BY 4.0) license.



Data publication + sensitive data + versioning

Plenary

Maitane Iturrate-Garcia

METAS

Funder Requirements

- Project proposal (JRP)
  - Open Science
  - Management of data + other outputs

Section B: Overview of the research

B1 Summary of the project

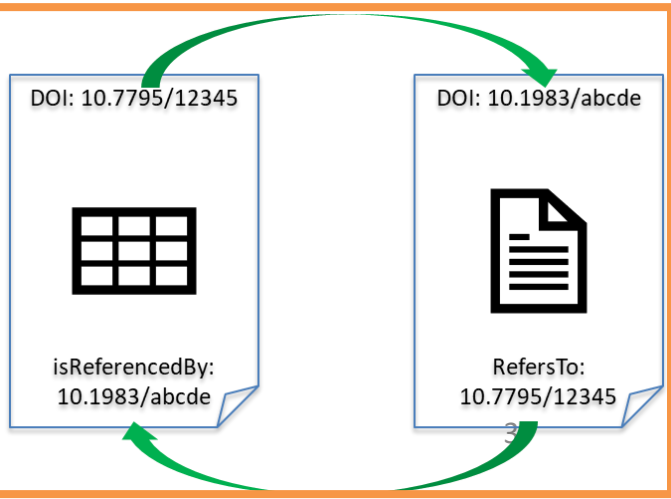
B2 Excellence

- B2.a Overview of the objectives
- B2.b List of deliverables
- B2.c Need for the project
- B2.d Progress beyond the state of the art
- B2.e Gender dimension
- B2.f Open science
- B2.g Research data management and management of other research outputs

B3 Potential outcomes and impact from the project

- Data management plan
  - FAIR principles; Security; Licenses; Responsibilities; Costs…

- Article & data publication
  - Funder acknowledgement
  - Metadata, identifiers…



EURAMET EPM requirements

EPM projects are “co-financed from the EU Horizon Europe Research and Innovation Programme and by the Participating States”.

Open science practices as per Regulation (EU) 2021/695 (Horizon Europe): rules for participation and dissemination

JRP proposals (max. 1 pages) in section B2.f (Open science):

- Early and open sharing of research outputs
- Providing open access to research outputs (publications covered in Annex 5 GA)
- Participation in open peer-review

“As open as possible, as closed as necessary”

EURAMET EPM requirements

“As open as possible, as closed as necessary”

All data needed to validate results in scientific publications **openly available** (no raw data)

By default but possible to restrict data sharing on a **case-by-case basis**

Examples of voluntary restrictions given by EURAMET:

- Third parties’ data (if they don’t agree to make them openly available)
- Data that disclose identity of a manufacturer
- Data compromising the protection of a participant(s) intellectual property

Personal and sensitive data

Data protection is a fundamental right under EU law: General Data Protection Directive (GDPR), Law Enforcement Directive (LED) and Data Protection Regulation for EU institutions, bodies, offices and agencies (EUDPR).

Personal data: data that can be used to identify a person

Sensitive data: a special type of personal data that requires extra precautions (e.g., genetic data, data relating to health, ethnicity, religion, criminal proceedings…)

Personal/sensitive data can be stored, edited and shared considering certain aspects

Personal and sensitive data

At the beginning of the project:

- Organize secure data storage (get in touch with IT) – only authorized access
- Prepare consent forms (following legal and ethical guidelines and requirements)

Before publishing:

- Make sure you have participants' consent
- Data anonymization (data obfuscation, data pseudonymization…)
- Select a licence to regulate how data can be reused
- Pay attention to backup strategies (cloud options data protection-friendly?)

Data anonymization

Data is hidden or transformed (e.g. aggregated, regrouped or deleted) in a way is not possible to reidentify individuals and/or to have unauthorized access.

Generalization

	Name	Age	Salary	Location
0	Alice	28	55000	123 Main St, New York
1	Bob	45	85000	456 Elm St, Los Angeles
2	Charlie	33	45000	789 Oak St, Chicago
3	David	19	30000	234 Maple St, Houston
4	Eve	61	90000	567 Pine St, Miami

	Name	Age	Salary	Location
0	Alice	26-35	55000	New York
1	Bob	36-50	85000	Los Angeles
2	Charlie	26-35	45000	Chicago
3	David	0-25	30000	Houston
4	Eve	51+	90000	Miami

Pseudonymization

	Name	Email
0	Alice	alice@example.com
1	Bob	bob@example.com
2	Charlie	charlie@example.com
3	David	david@example.com
4	Eve	eve@example.com

	Pseudonymized_Name	Pseudonymized_Email
0	3BH0HXGW66	787LRWFOY8
1	L8CYVHMKEQ	6S6PMDONFJ
2	MKXH8SIKOR	DXA8PIILRA
3	IZX98513Z8	RAQ8T0NPNV
4	D9HM15XORG	8BSZ8VT2S1

Source: GeeksforGeeks

Masking

	Name	SSN	Phone	Credit Card
0	Alice	123-45-6789	555-1234	4111-1111-1111-1111
1	Bob	987-65-4321	555-5678	5500-0000-0000-0004
2	Charlie	456-78-1234	555-8765	3400-0000-0000-009
3	David	321-54-9876	555-4321	3000-0000-0000-04
4	Eve	654-32-1987	555-6789	6011-0000-0000-0004

	Name	SSN	Phone	Credit Card
0	Alice	XXX-XX-6789	555-XXX	XXXX-XXXX-XXXX-1111
1	Bob	XXX-XX-4321	555-XXX	XXXX-XXXX-XXXX-0004
2	Charlie	XXX-XX-1234	555-XXX	XXXX-XXXX-XXXX--009
3	David	XXX-XX-9876	555-XXX	XXXX-XXXX-XXXX-0-04
4	Eve	XXX-XX-1987	555-XXX	XXXX-XXXX-XXXX-0004

Source: DataCamp

Data versioning

Data integrity and reliability, reproducibility, troubleshooting, traceability, provenance

Two types

- Versioning of data being generated (early in the project)
- Versioning of published dataset (towards the end of the project)

Backup vs. Versioning

Data backup: creating copies of data for recovery purposes

Data versioning: tracking and managing dataset changes over time (metadata!!)

Challenging (e.g. large datasets, pipeline integration)

Data versioning – Type 1

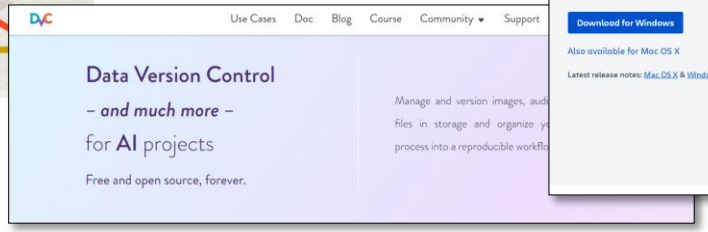
Best practices

- Right **versioning strategy** depending on specific needs and constraints (version numbering schemes, when to create new versions, version storage…)
- Effective **metadata** management (metadata describing changes, context, …)
- Clear **policies/guidelines** within consortium (consider partners' institution policies!!)
- **Automatic** version control (existing specialized tools; e.g. Git-based solutions)

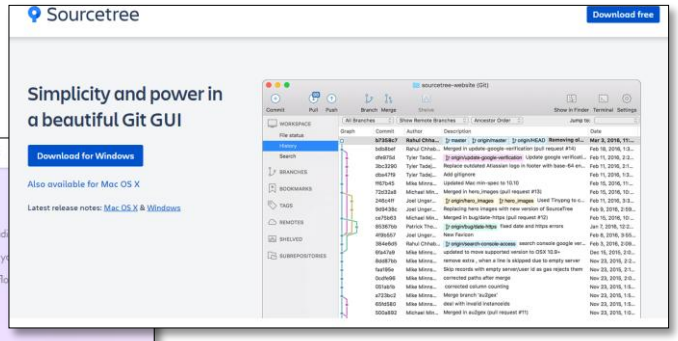
https://git-scm.com

https://github.com

https://dvc.org



https://www.sourcetreeapp.com



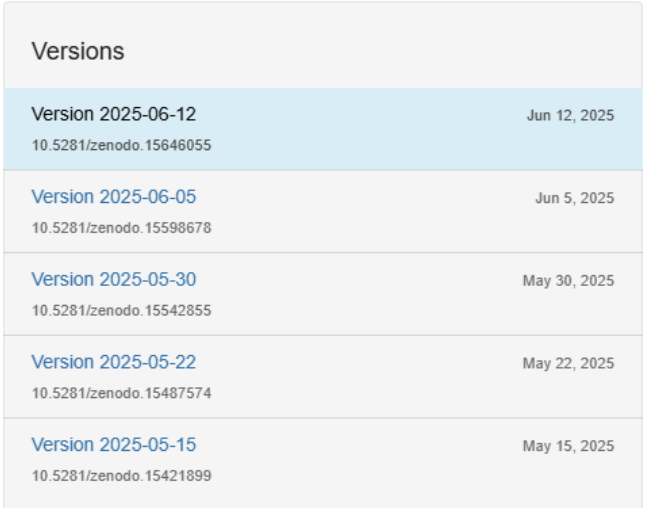
Data versioning – Type 2

Datasets published in open repositories

- Repository metadata can be modified after publication
- Uploaded files cannot be modified!

To “modify” published datasets, new versions upload to repository

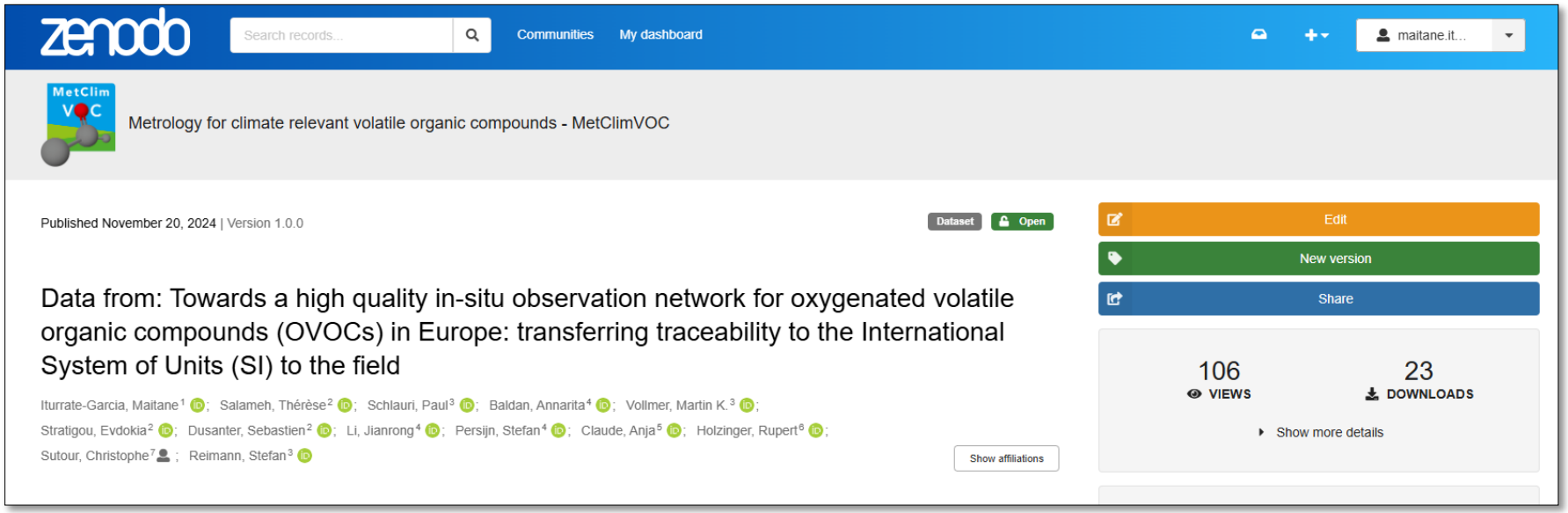
New versions associated with the original data (a kind of common DOI comprising all versions + an individual version DOI)



Source: https://zenodo.org/records/15646055

Data versioning – Type 2

https://zenodo.org/records/14178374



Data publication

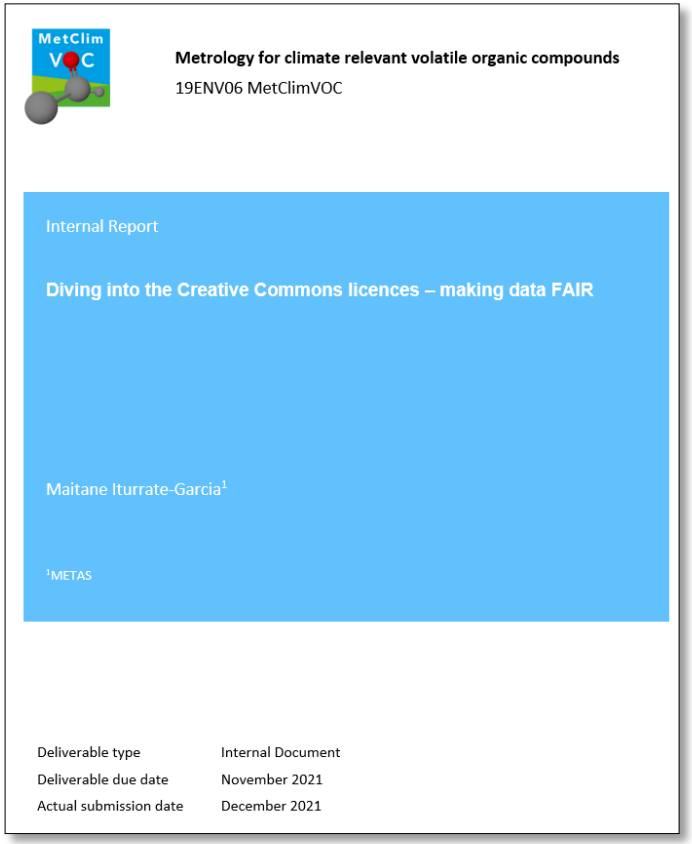
Several aspects to consider at the beginning of the project (part of RDM)

- Early discussions within the consortium (agreement): open repository, licence type, datasets to be published/restrictions, responsible person/small group for data curation, data curation policies…
- Consortium training (licensing, metadata, data versioning…)
- Project flowchart and checklist for data publication covering the main processes: dataset preparation for publication, metadata creation, long-term access, licensing, data citation…

Data publication – a real-case

A real-case experience (EMPIR 19ENV06)

Consortium training/early discussions



MetClim VOC

Metrology for climate relevant volatile organic compounds

19ENV06 MetClimVOC

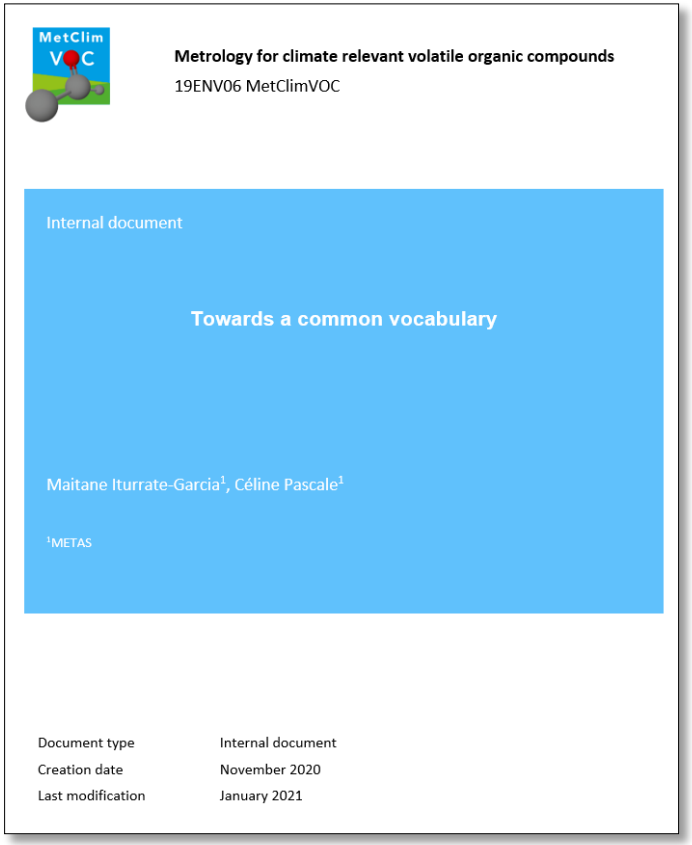
Internal Report

Diving into the Creative Commons licences – making data FAIR

Maitane Iturrate-Garcia[1]

[1]METAS

Deliverable type	Internal Document
Deliverable due date	November 2021
Actual submission date	December 2021



MetClim VOC

Metrology for climate relevant volatile organic compounds

19ENV06 MetClimVOC

Internal document

Towards a common vocabulary

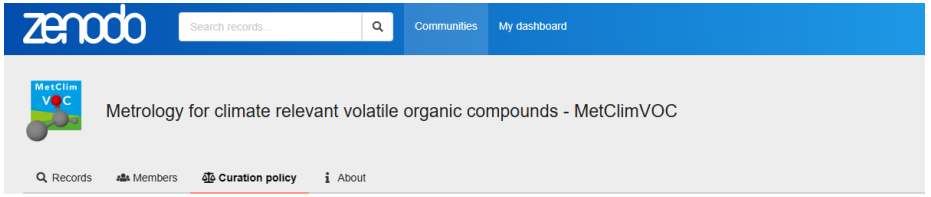
Maitane Iturrate-Garcia[1], Céline Pascale[1]

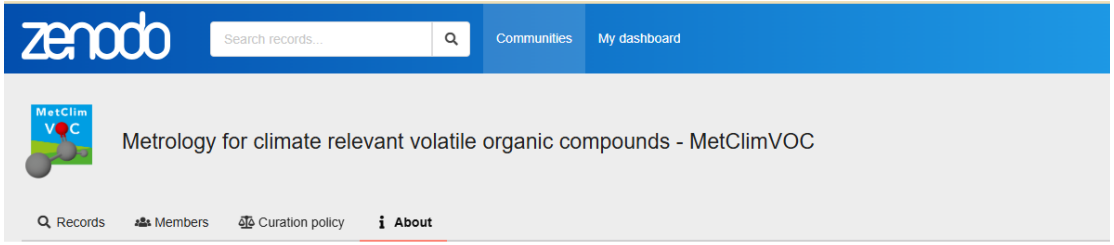
[1]METAS

Document type	Internal document
Creation date	November 2020
Last modification	January 2021

Data publication – a real-case

Early discussions/decisions





A community to share datasets, papers and guidelines contributing to improve the traceability, accuracy and comparability of measurements of atmospheric volatile organic compounds (VOCs) relevant for climate. The main goal of the community, created by the European Project EURAMET-EMPIR 19ENV06 MetClimVOC (https://www.metclimvoc.eu), is to provide open access FAIR data with focus on metrological aspects of atmospheric VOCs measurements.

The collection includes mainly data generated in the framework of the project MetClimVOC from measurements, calibrations, comparisons and validations. In addition, datasets originated out of the project but contributing clearly to improve the metrology of VOC measurements are accepted (see curation policy).

The collection is suitable for use by research groups working on the following topics: climate change, atmospheric chemistry, reactive gases, VOCs and spectrometry, as well as for atmospheric monitoring networks and bodies, and manufacturers of instruments, gas handling devices, gas cylinders and VOCs dynamic methods (diffusion, permeation…).

1. If an uploaded document and/or dataset has an existing DOI, it will be kept. A Zenodo DOI will be attributed to documents and/or datasets without existing DOI.
2. Items with Open Access remain Open Access.
3. Items must not contain any copyright restrictions.
4. All files should be able to be opened without any passcode restrictions.
5. No Personal Health Information or sensitive data can be included.
6. All information needs to be in English.
7. Datasets will be licensed under a Creative Commons Attribution 4.0 (CC BY 4.0) or a Creative Commons Attribution Share Alike 4.0 (CC BY-SA 4.0).
8. Datasets should include descriptive names for columns and rows.
9. A README file must be provided together with the dataset including at least the following information:
   - Title: it should be descriptive of the dataset
   - Author(s): name, institutional affiliation, email address and/or ORCID
   - Abstract: short description of the dataset
   - Methods: any methodological information (i.e. analyser, compounds…), including calibration and traceability information, which may help others to understand how the dataset was generated.
   - Funding information (e.g. in the case of datasets generated in the framework of the project MetClimVOC, this information will include the terms "EMPIR", "European Union (EU)" and "Horizon 2020" as required in article 29.2 of the EMPIR Grant Agreement)
   - Related works: e.g. publication DOI, other datasets, code…

The community curator will use the information included in the readme file to generate the metadata of the dataset. The metadata will be human and machine-readable (e.g. JSON).

To upload datasets and publications, please contact maitane.iturrate@metas.ch or celine.pascale@metas.ch

Data publication – a real-case

Data publication process

- During manuscript drafting
- Before submitting manuscript
- After manuscript acceptance
- Dataset uploading (only for data curator)

MetClim VOC

Metrology for climate relevant volatile organic compounds

19ENV06 MetClimVOC

Report on research data and metadata

Guidelines on the preparation of datasets and metadata to ensure the FAIRness of MetClimVOC data

Maitane Iturrate-Garcia[1]

[1]METAS

Deliverable type	Internal Report
Creation date	August 2022
Last modification	January 2023

Data publication – a real-case

(Before) manuscript drafting

- Data errors
- Data inconsistencies
- Duplicates
- Irrelevant data
- Sensitive data

Data publication – a real-case

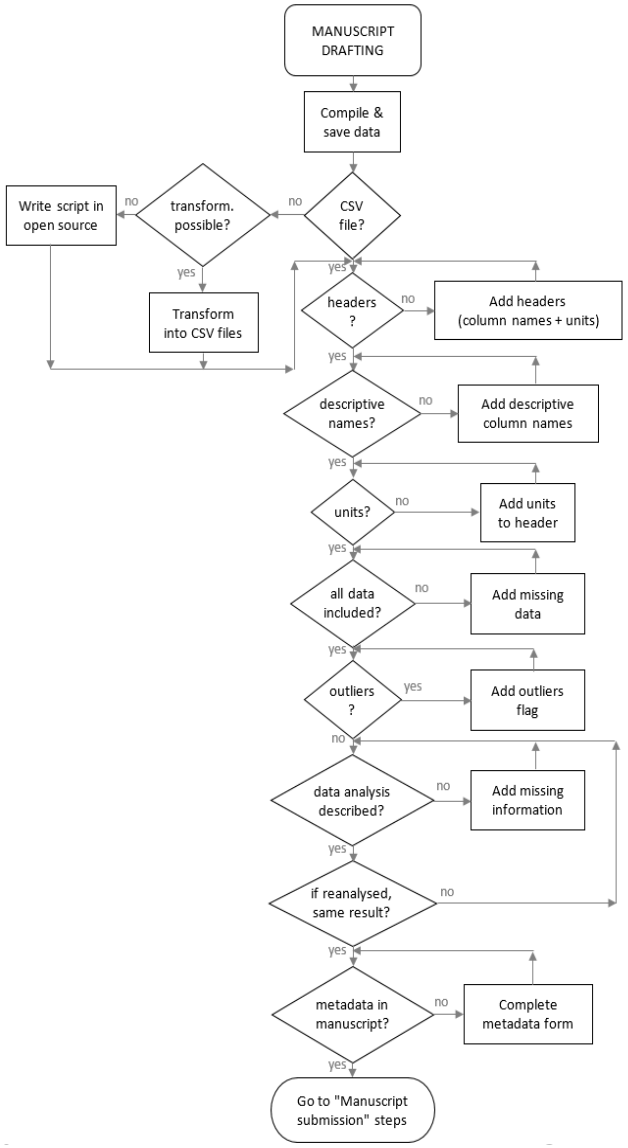
Manuscript drafting

Annex I: Checklist

Manuscript drafting

- ☐ All data are saved as CSV files.
- ☐ File names follow DMP specifications:

  yyyymmdd_19ENV06_MetClimVOC_*experiment01*_institution.csv.
- ☐ Descriptive headers are included in the files (column names including quantities and units).
- ☐ The requested metadata is available in the manuscript, dataset or (if needed) in the metadata form.
- ☐ The data analysis is described in detail indicating performed data transformations, outliers and software used for the analysis, including software version (this applies to programming packages also).
- ☐ All the questions included in Annex II – Manuscript drafting flowchart – can be positively answered.



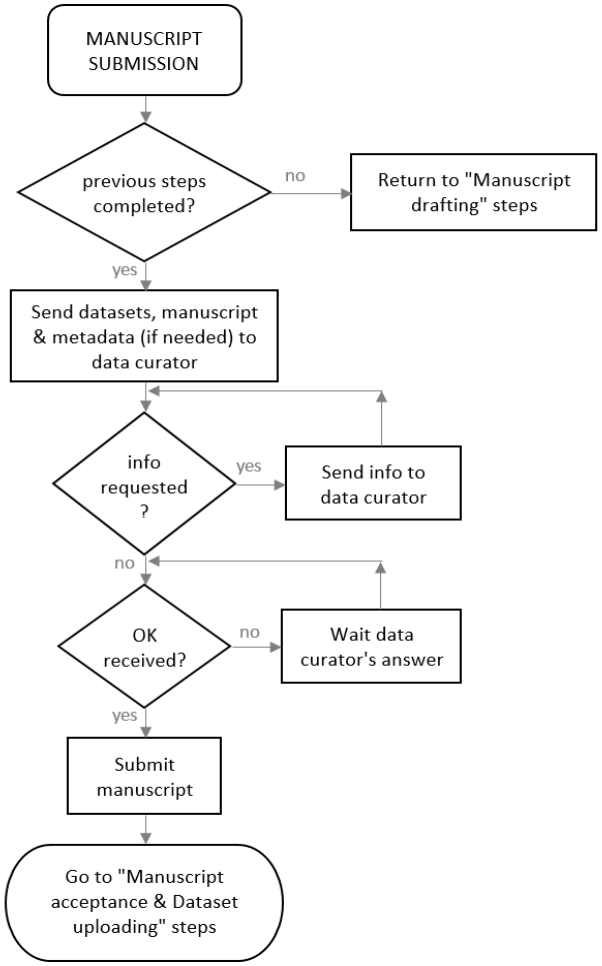
Data publication – a real-case

Manuscript submission

Annex I: Checklist

Manuscript submission

- ☐ Dataset, manuscript and metadata (if needed) were sent to the data curator.
- ☐ Data curator's OK was received.
- ☐ Manuscript submission.
- ☐ Manuscript review finished; manuscript accepted.



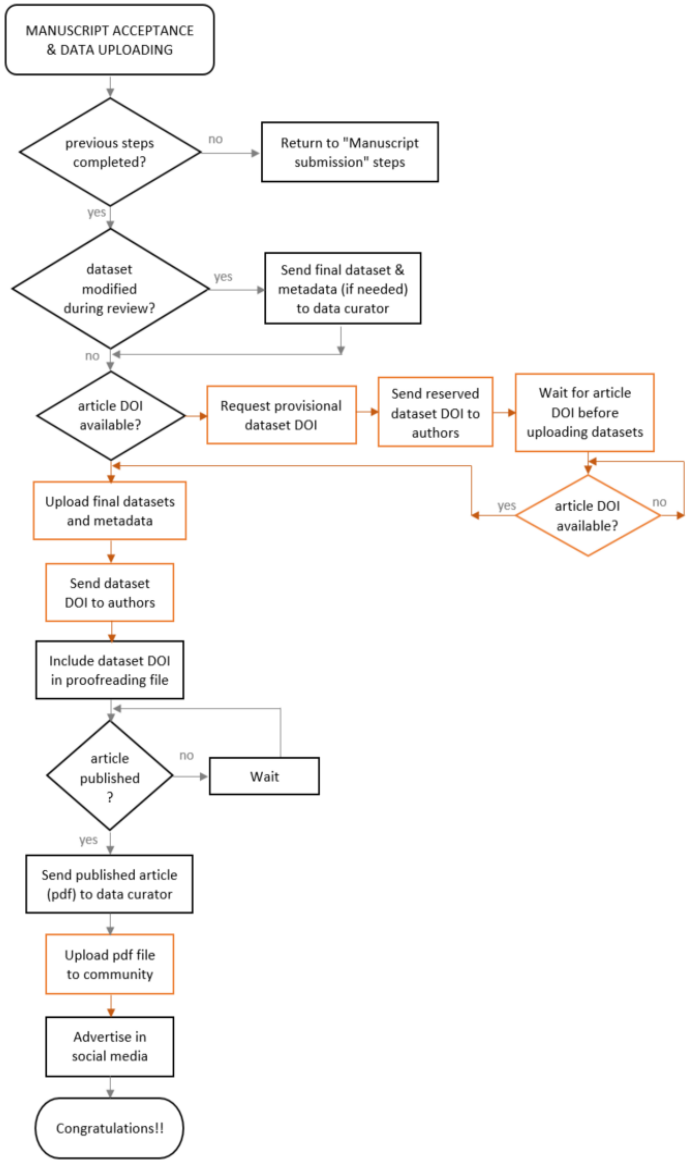
Data publication – a real-case

Dataset uploading

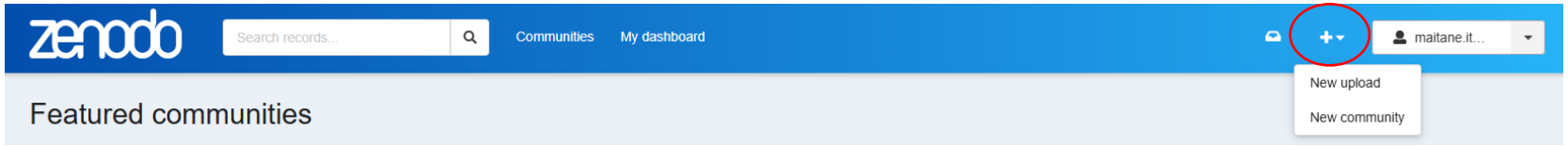
Annex I: Checklist

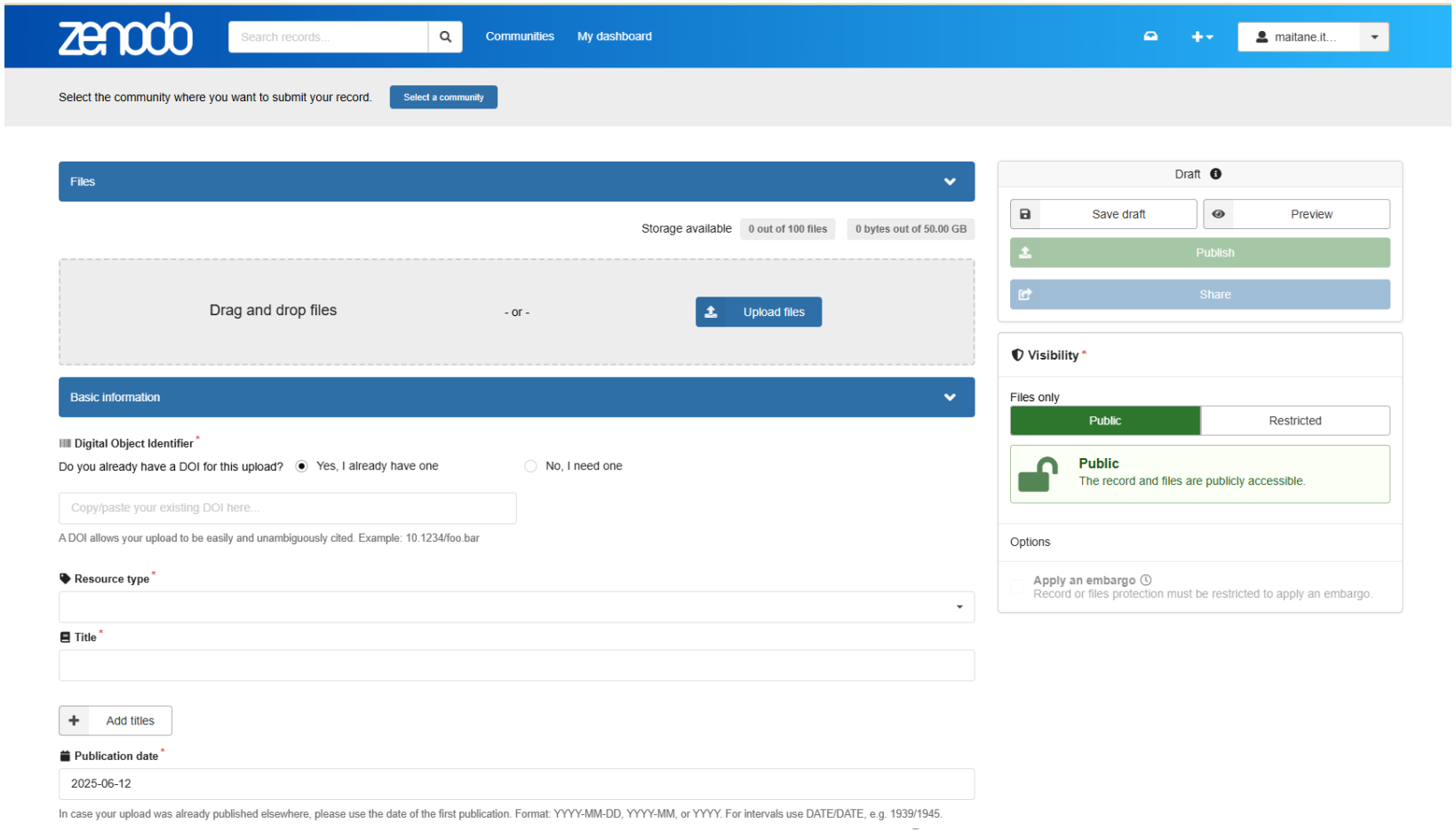
Dataset uploading

- ☐ Final dataset (if modified during review process) sent to data curator.
- ☐ (Only data curator) data files format and structure checked: correct and metadata completed.
- ☐ (Only data curator) DOI for dataset ("reservation) requested if accepted article DOI is not available; information sent to authors.
- ☐ (Only data curator) datasets uploaded – only after the article DOI is available.
- ☐ Final pdf file sent to the data curator (only after the proofreading process is finished and the article published).
- ☐ (Only data curator) Article pdf file uploaded to the community.
- ☐ (Authors + data curator) published article and uploaded datasets advertised in the personal and project social media.
- ☐ (Only data curator) dataset and article information sent to EURAMET repository.

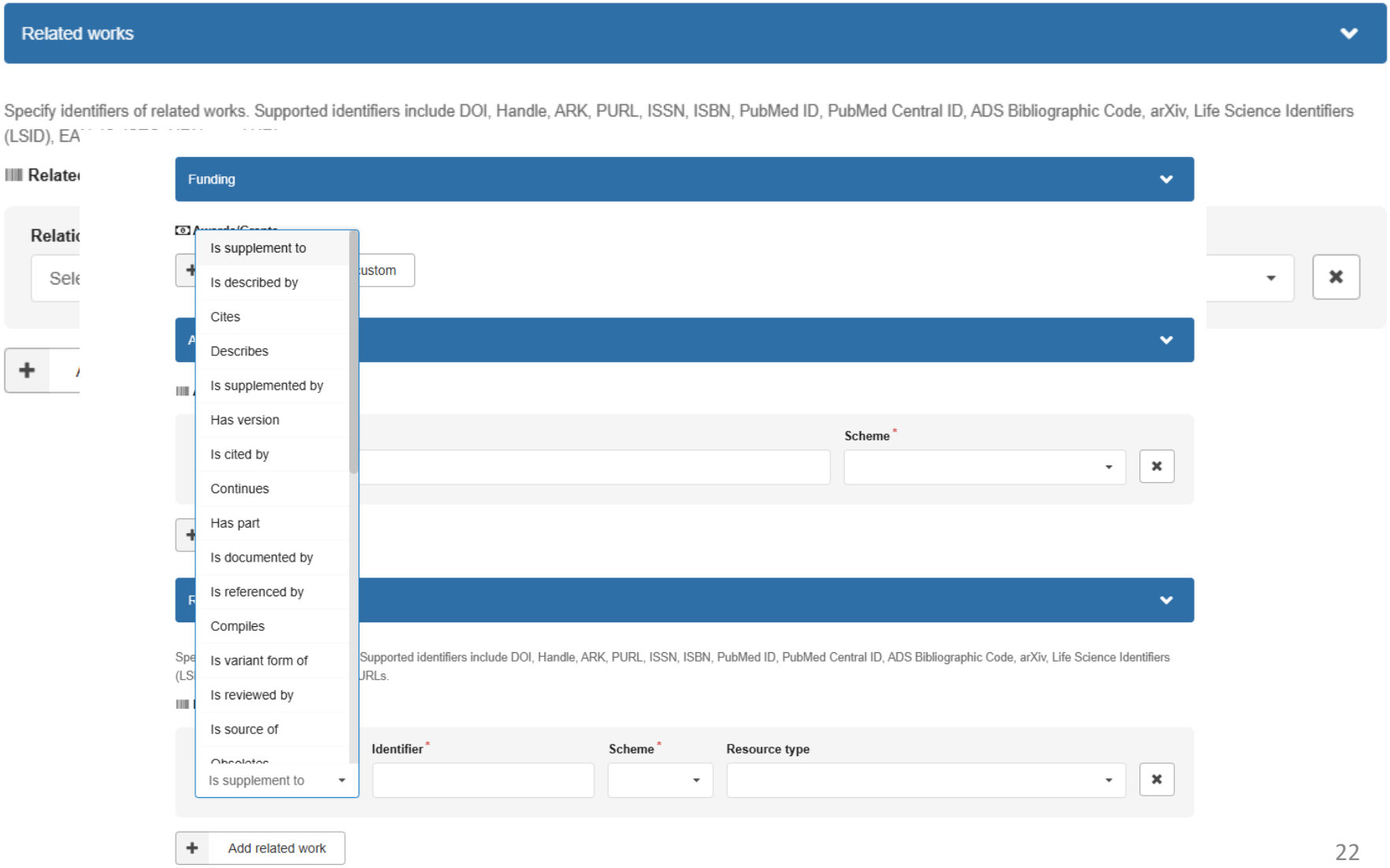


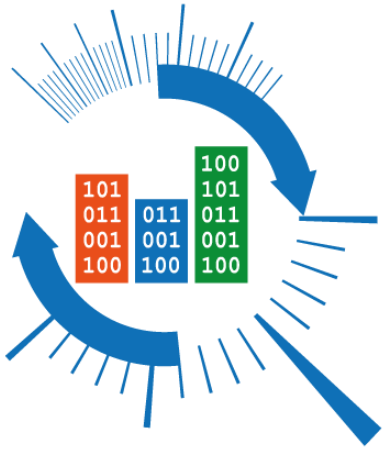
Data publication – a real-case





Data publication – a real-case



Take-home messages

- RDM requires (initial) great amount of resources, but a good RDM will save you many headaches
- Don’t underestimate the power of communication
- You are not alone!!