
DTU

DOI: 10.5281/zenodo.3999354

25/08/2020

Webinar: ASK ME ANYTHING

Nikola Vasiljevic

Introduction to FAIR and R5 principles

Disclaimer

- *This presentation contains personal opinions which not necessarily reflect the DTU Wind Energy or Research Data Alliance politics.*

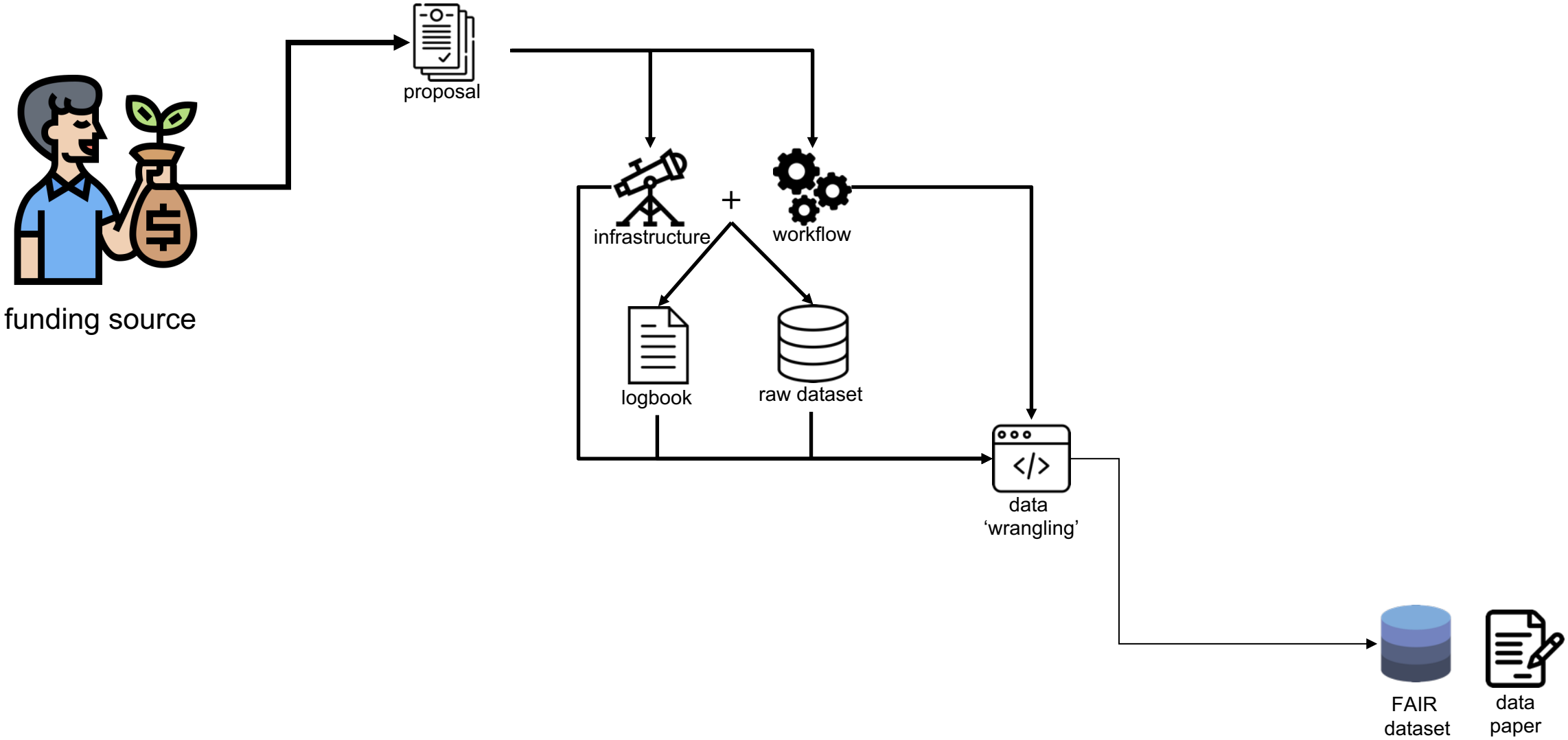
- *Most of resources in slides are sourced from : https://like-itn-digitalization.readthedocs.io/en/latest/*

Table of Contents

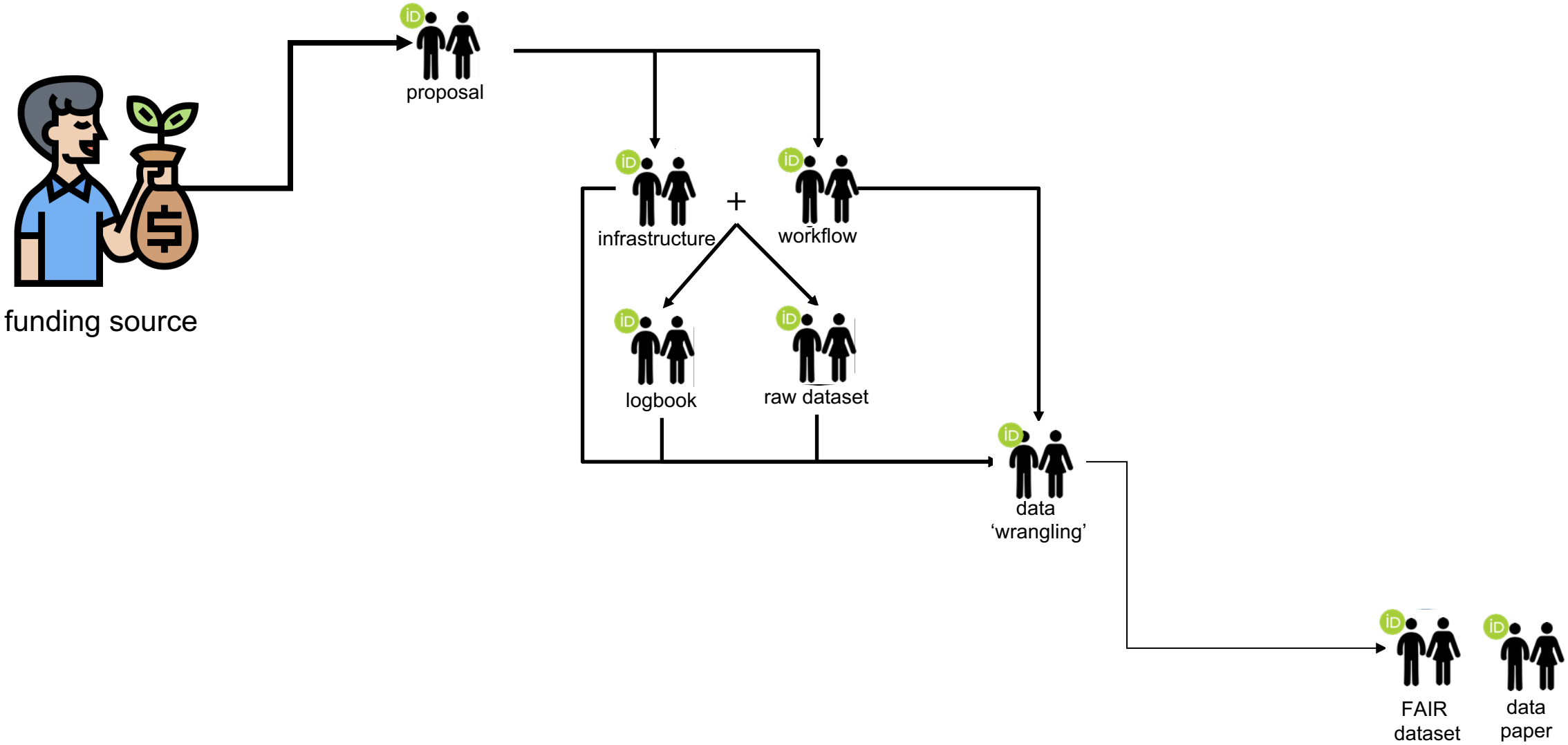
01

Research Project Lifecyle and Open Science

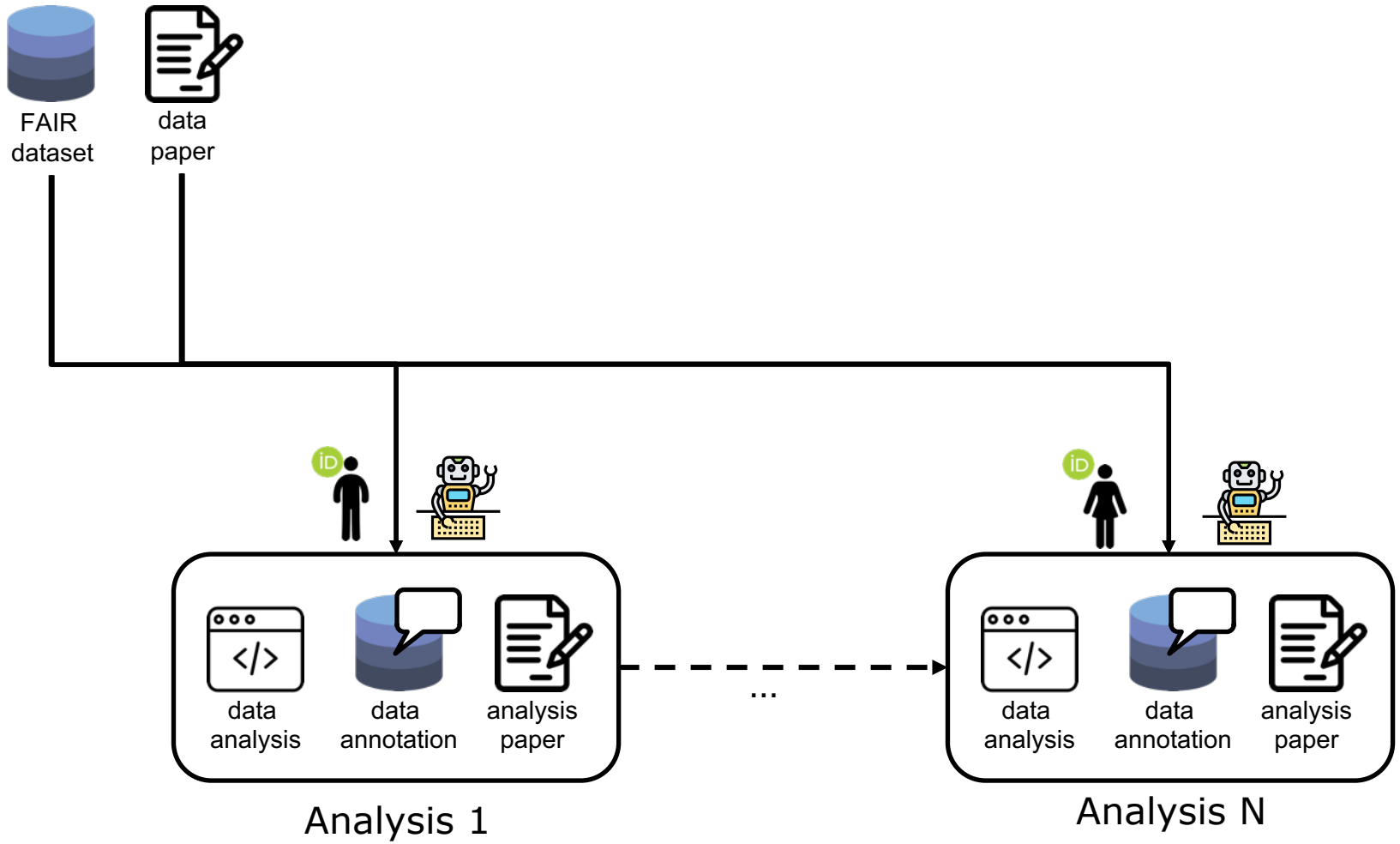
Blueprint of idealized project lifecycle



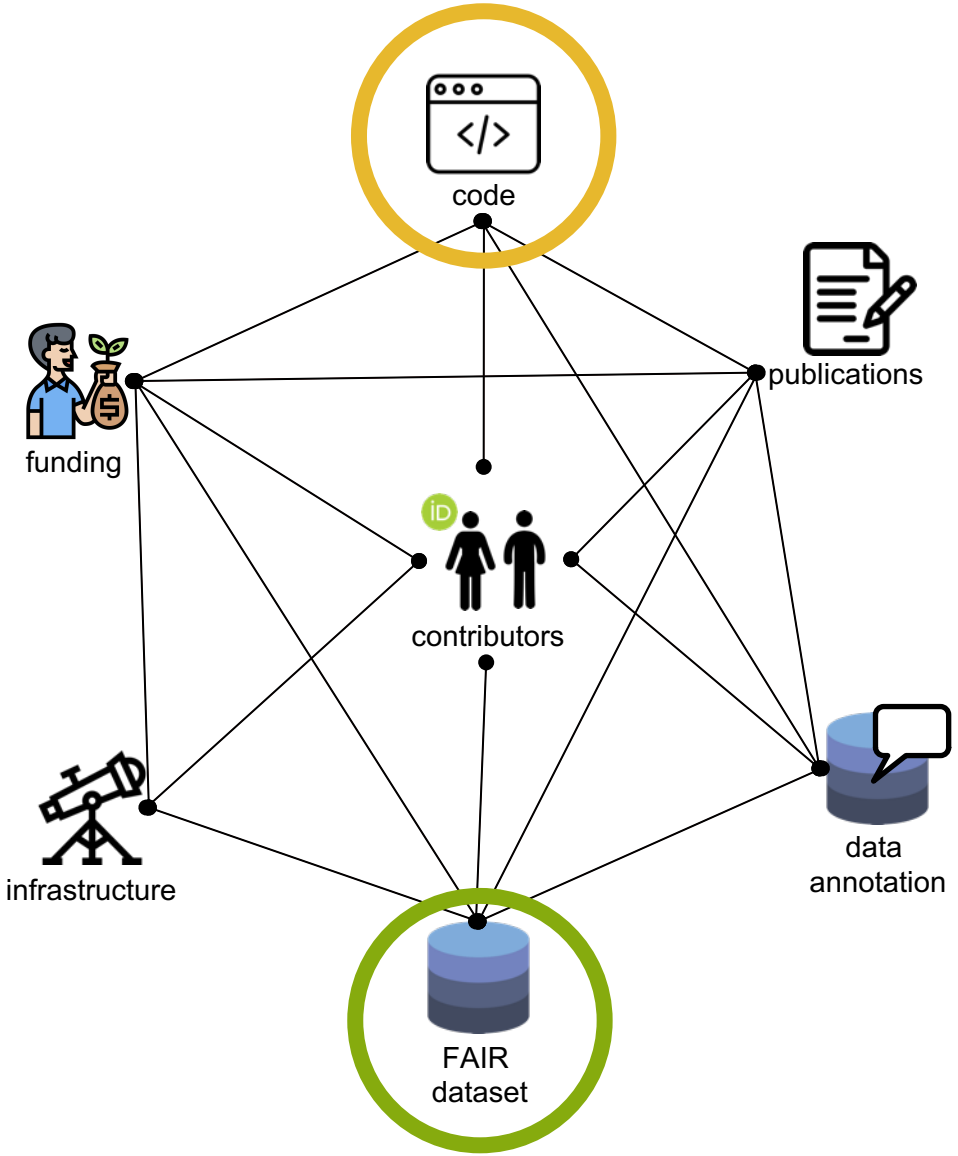
Blueprint of idealized project lifecycle



Blueprint of idealized project lifecycle



Relations



Full chain of custody

Researcher at center

Remember!

If you don’t have ORCID ID, make one ASAP!

It is your unique ID that you will link all your work to!

Typically people are publishing only

- Traditional journal articles

But you can publish much more

- Data
- Data papers
- Code
- Code papers
- Traditional journal articles

How you publish things matters!

✓	Open access to abstract
✗	Open access to paper
✗	Open access to lab notebooks
✗	Open access to underlying data
✗	Open access to underlying code
✗	Version control of data and code

‘Traditional’ approach

Source: Elsevier logo Wikipedia

How you publish things matters!

☑	Open access to abstract
☑	Open access to paper
☑	Open access to lab notebooks
☑	Open access to underlying data
☑	Open access to underlying code
☑	Version control of data and code

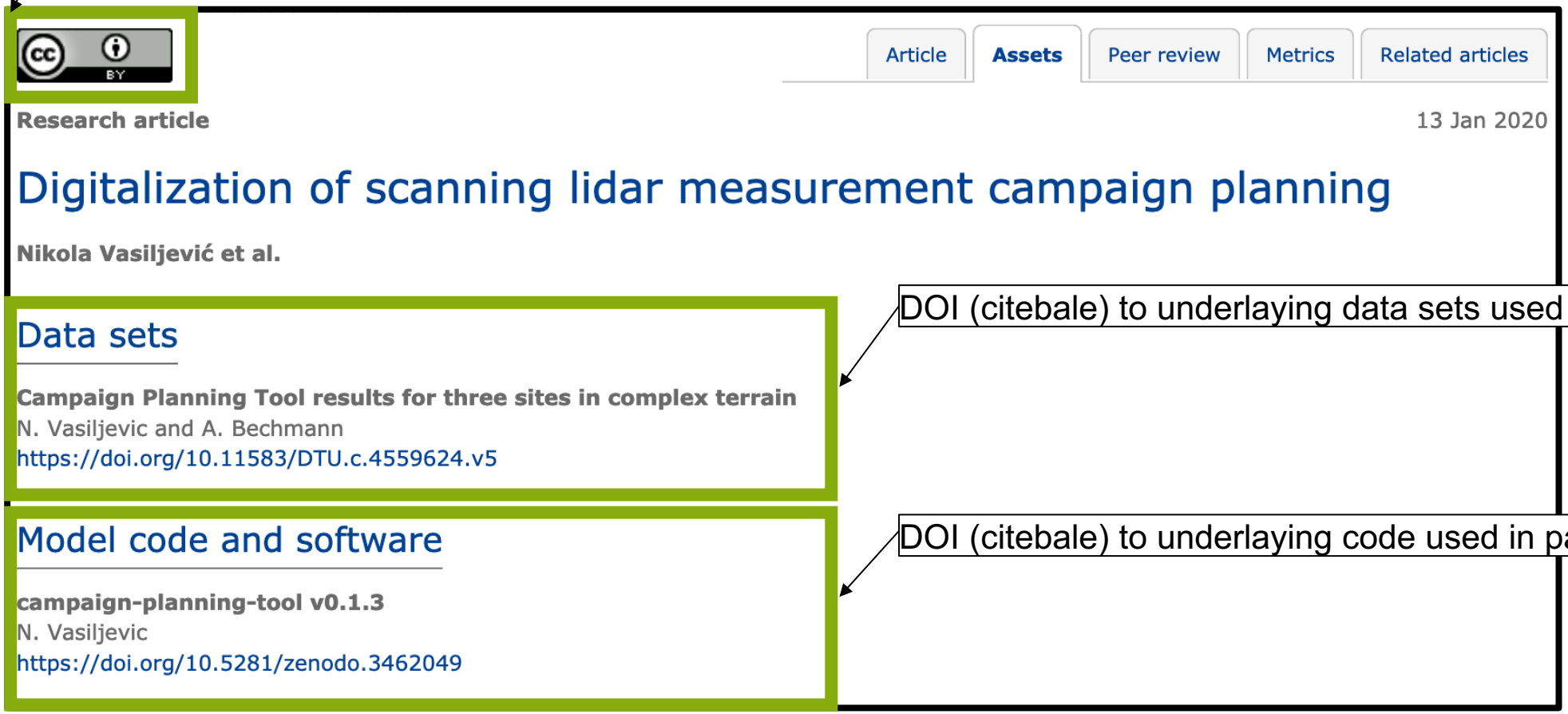
Open Science approach

Few examples of good practice

- https://wes.copernicus.org/articles/5/73/2020/
- https://wes.copernicus.org/articles/5/1059/2020/#bib1.bibx16
- https://amt.copernicus.org/preprints/amt-2020-321/

Clear usage license (Creative Commons Attribution)

- https://wes.copernicus.org/articles/5/73/2020/



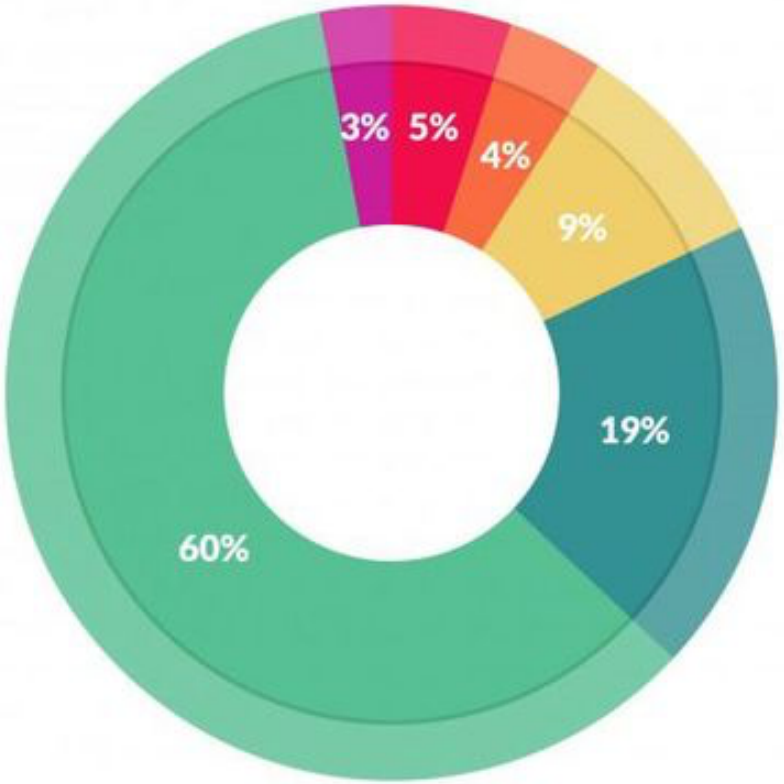
DOI (citebale) to underlaying data sets used in paper

DOI (citebale) to underlaying code used in paper

02

FAIR Data principles

How much time we spend analysing data?

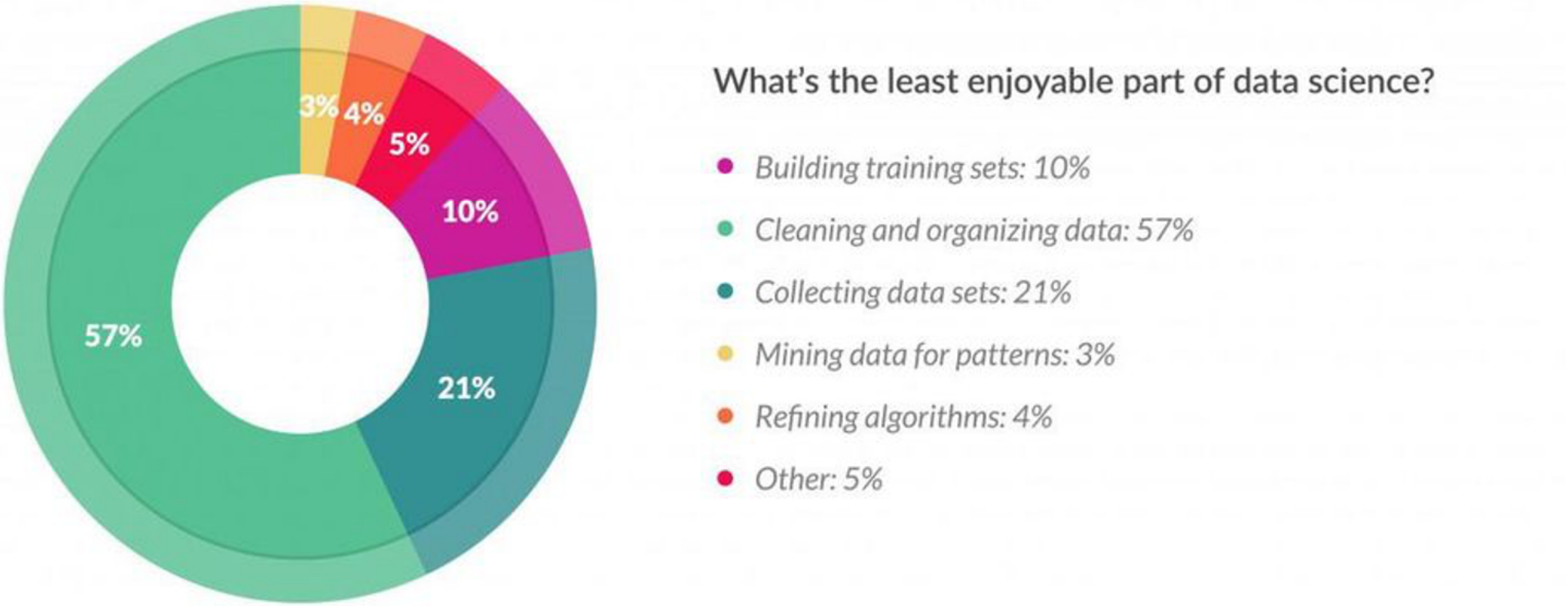


What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: Forbes

What we don’t like to do?

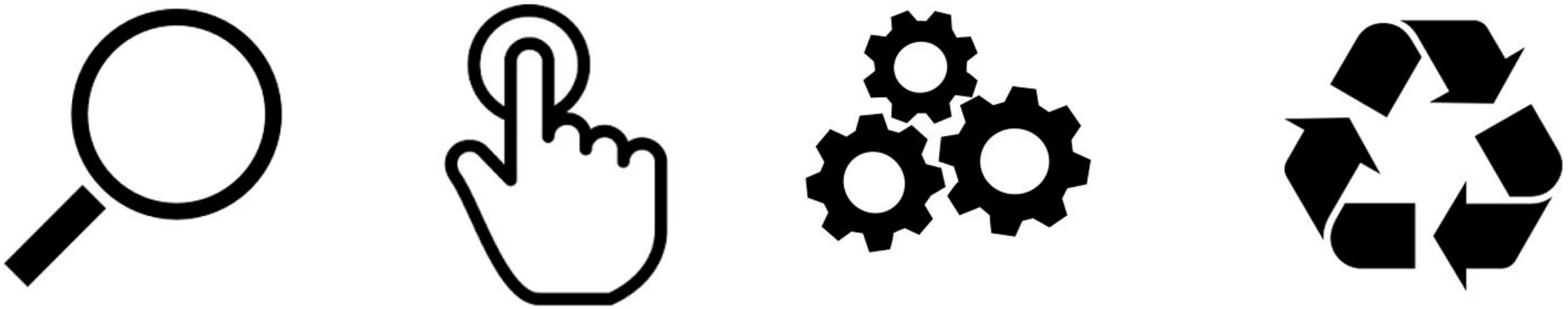


Source: Forbes

If we don’t spend time handling data at the moment of their creation we will waste 80% of resources anytime we or anyone else need to use them (again).

Data engineering is not perceived as ‘cool’ activity comapring to data analytics, however it has much more lasting impact then trendy data analytics methods.

Findable Accessible Interoperable Reusable



Source: wikimedia

FAIR principles

- In 2016 a divers range of stakeholders (academia, industry, funding agencies, scholarly publishers) designed and jointly endorsed a concise set of principles to improve the infrastructure supporting the reuse of data

- These principles are known as the FAIR Data Principles

- FAIR stands for:
  (1) Findable
  (2) Accessible
  (3) Interoperable
  (4) Reusable

- These principles are intended to improve our data management and stewardship

- Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting data reuse by individuals

FAIR principles

F1. **Metadata** and **data** are assigned a globally unique and persistent identifier (**PID**)
F2. Data are described with rich metadata (further defined by R1 principle below)
F3. Metadata clearly and explicitly include the identifier of the data they describe
F4. Metadata and data are registered or indexed in a searchable resource

A1. Metadata and data are retrievable by their identifier using a standardized 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

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

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

FAIR principles

Technology

Social

F1. Metadata and data are assigned a globally unique and persistent identifier (PID)
F2. Data are described with rich metadata (further defined by R1 principle below)
F3. Metadata clearly and explicitly include the identifier of the data they describe
F4. Metadata and data are registered or indexed in a searchable resource

A1. Metadata and data are retrievable by their identifier using a standardized 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

I1. Metadata and data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2. Metadata and data use vocabularies that follow FAIR principles
I3. Metadata and data include qualified references to other metadata and data

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

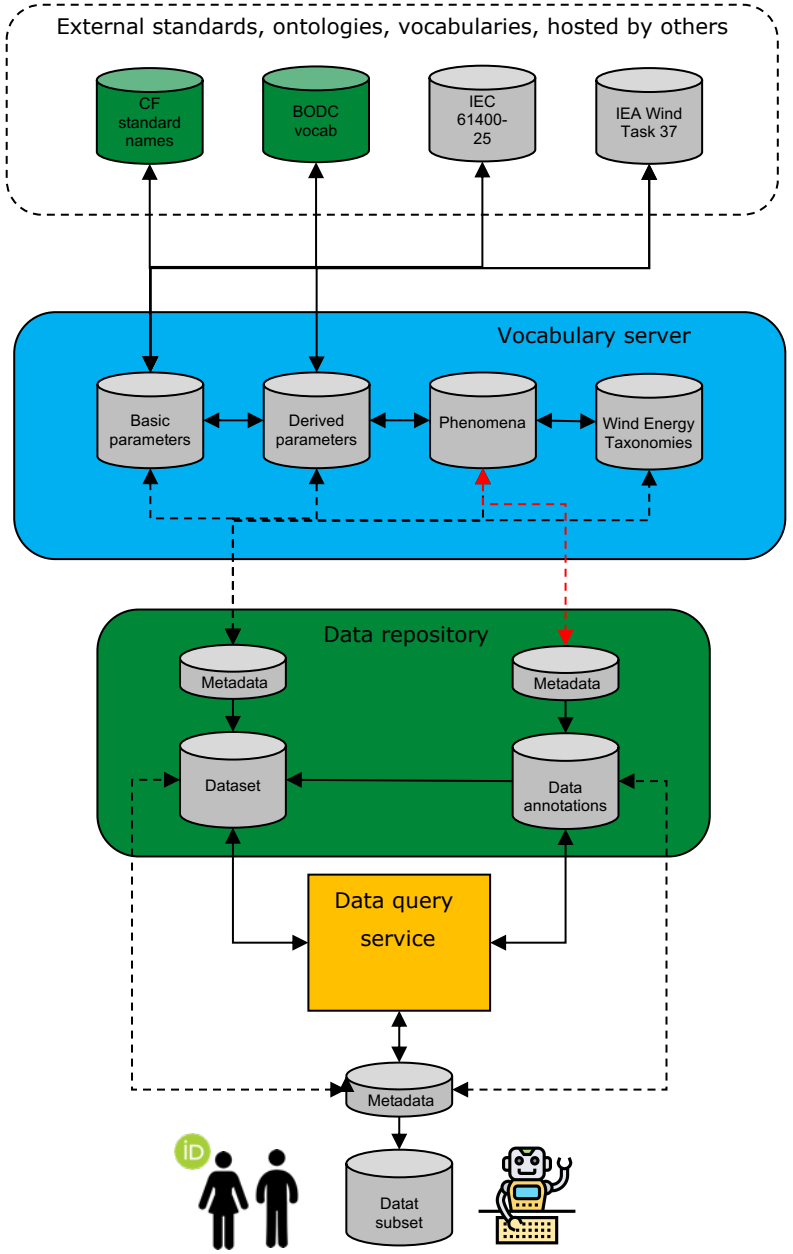
Source Erik Schultes, GO-FAIR foundation

Well, that's just fine and dandy—but what should I do?

‘Simple’ use-case

- Identify existing domain-specific controlled vocabularies which contain definition of parameters your data includes, search platform https://fairsharing.org/ for vocabularies from your domain (they call it terminology)
- Make sure that you always confirm to controlled vocabularies when you name parameters you are recording
- Identify existing domain-specific metadata templates that are used to report data, once again https://fairsharing.org/ could be a good starting point
- If metadata templates don’t exist, you will have to define them yourself (you could use data cite dataset metadata schema as an inspiration or one from Marinet2 project)
- Encode metadata template to open format (e.g., YAML, JSON, RDF, etc.)
- Select data standard which is widely used in your domain, remember it should be open standard.
- Publish your data and metadata in **Zenodo** or your university or domain specific data publishing platform

Advance use-case



Source: https://zenodo.org/record/3865225

03

R^5 Principles

R^5 principles

- Benureau and Rougier in [their work from 2018] defined five characteristics that a scientific code should possess in order to be regarded as **a first-class research output**

- These five characteristics translate into principles that those who write scientific code should follow

- In short, a scientific code should be:
  1. Re-runnable (R^1)
  2. Repeatable (R^2)
  3. Reproducible (R^3)
  4. Reusable (R^4)
  5. Replicable (R^5)

Re-runnable (R^1)

The code should be re-runnable (i.e., re-executable).

Repeatable (R^2)

The code should produce the same results every time it is executed.

Reproducible (R^3)

The code should allow those who use it to reobtain the published results.

Reusable (R^4)

The code should be easy to use, understand and modify.

Replicable (R^5)

A clear an unambiguous algorithmic description of the code should available as a reference allowing the code replication in other programming languages than one chosen for an initial implementation.

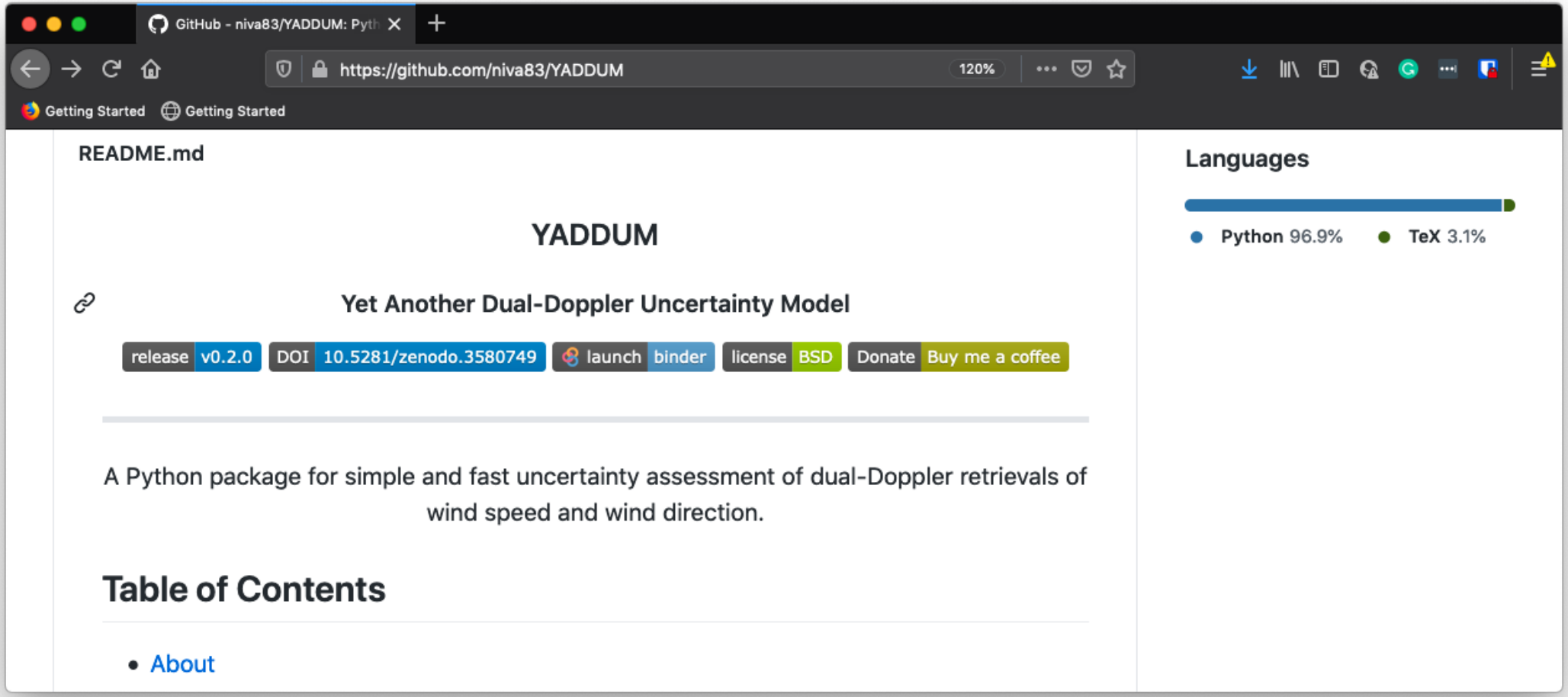
Recommendations for scientific code

- Select programming language and conform to that language standards and community endorsed code style
- Document your code
- Use Git to version control your code (preferably use GitHub see why below)
- Like with data, always select appropriate license (e.g., BSD, MIT, etc.)
- If you are going to use Github, connect your Github profile to your Zenodo profile and get for free ability to:
  - persist your code
  - version control your code at two location (Zenodo + Github)
  - automatically publish your code with every new version
  - make your code and all its versions citeable
- Publish description of your code as a short-paper in Journal of Open Science Software (it requires you to have your code on Github)

Some examples

- https://github.com/niva83/YADDUM
- https://github.com/niva83/mocalum
- https://github.com/niva83/campaign-planning-tool

- https://github.com/niva83/YADDUM



Version control of your code

DOI for citing

Run your code in browser without a need to install it

Usage license

Summary

- Comparing to ‘traditional’ academic practices **being a researcher in the age of digitalization** and Open Science requires some extra work:
  – your data should be in good shape (metadata and data standards, etc.)
  – your code should look more as **professional-ware** instead of **professor-ware**

- However, you will reap benefits of your extra and meticulous work:
  – immediately, by having more and diverse research outputs published
  – in future, by not wasting time understanding what you did a month or years ago
  – opens up more career paths than only being a researcher (e.g., programmer, data engineer, etc.)

Thank you for your attention
Questions?

niva@dtu.dk