Before you go into the field: preparing how to collect data

BIODIVERSITY.AQ

Biodiversity data from the field to research

Anton Van de Putte
SCAR Antarctic Biodiversity Portal

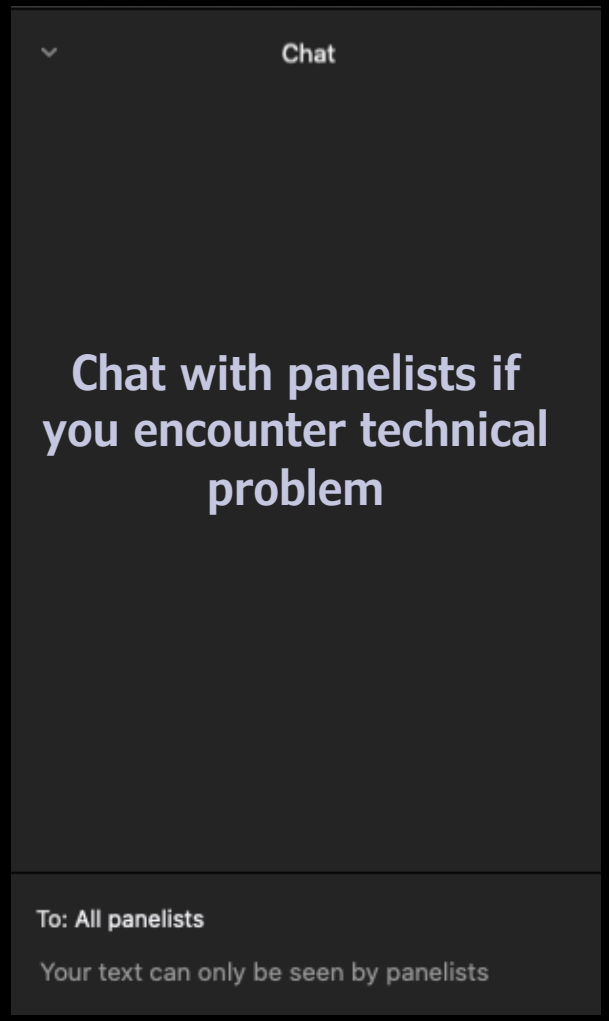
We are recording this seminar

Let us know if you prefer not to be recorded.

Code of Conduct

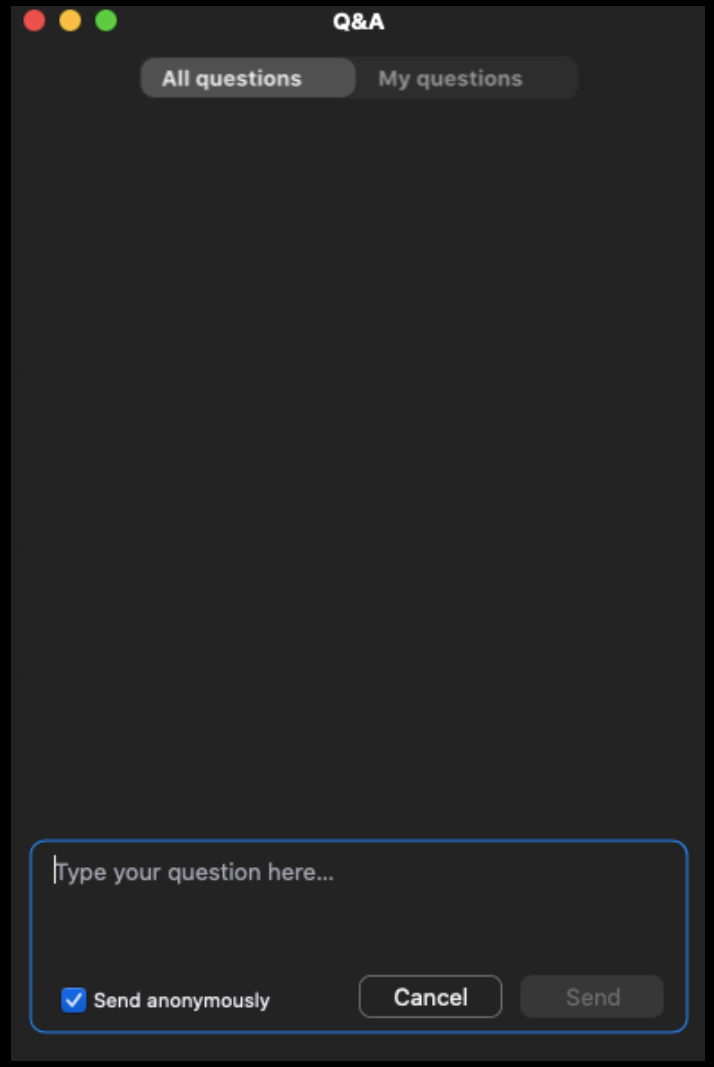
- Be respectful
- We will follow the principles of the rOpenSci Code of Conduct
  - https://ropensci.org/code-of-conduct/

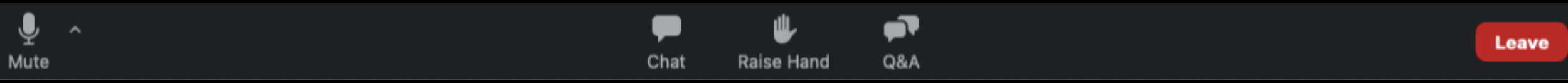
Using Zoom in this webinar



Chat with panelists if you encounter technical problem

Ask questions using Q&A feature or raise your hand





This course

- **Before you go into the field: preparing how to collect data**
- **Using a template to structure data: practical tips and tricks**
- **How to Quality Control your data**
- **Collecting public data: where and how to get open data**

These webinars wont cover everything

- Discover more based on the links that we provide
- Schedule a session to discuss directly
  - https://doodle.com/meetme/qc/XxYYpJwmbG

SCAR Antarctic Biodiversity Portal

- Advice & support
- Finding data
- Data publication
- Development of data tools

museum

LifeWatch BELGIUM

Belgian Science Policy Office belspo

SCADM

SCAR

BIODIVERSITY.AQ

OBIS OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM

GBIF.org

There is help

- SCAR Antarctic Biodiversity Portal
- SCAR expert Group on Antarctic Biodiversity Informatics
- Standing Committee of Antarctic Data Managers
- Southern Ocean Observation System

About You

Tools for quality control (your) data

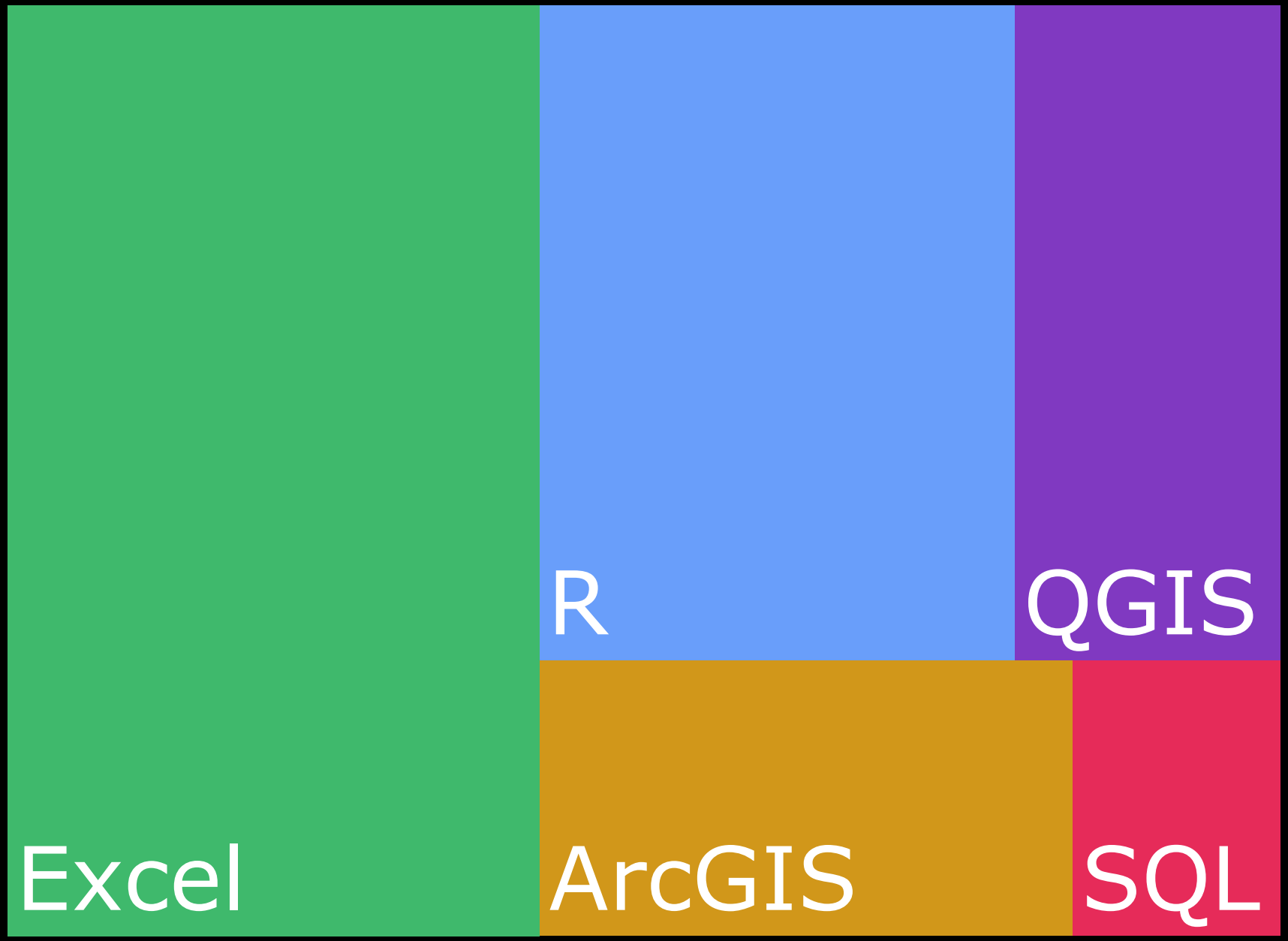
General introduction to biodiversity data management and field work preparation

Integrating your data with public data

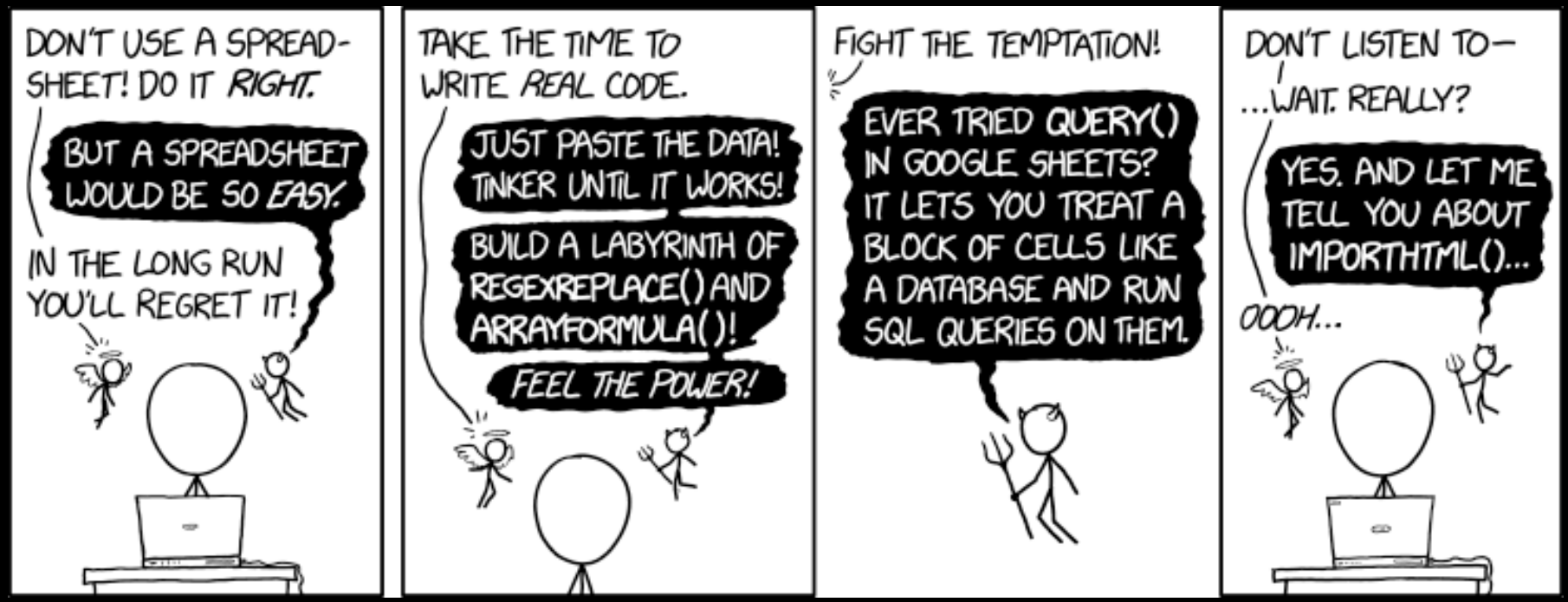
Templates for recording biodiversity data in the field

- visualization of biodiversity data
- Develop data management training materials
- Biodiversity data analysis using open source statistical softwares (R, Python)
- Preparation of tables for data analysis
- ensuring data is formatted to be accessible for other researchers
- Survey Opportunities

Software



CATCH-22



Don’t use Excel

Everybody uses Excel

Key Principles

Antarctic Treaty

- Antarctica shall be used for peaceful purposes only
  Art. I
- Freedom of scientific investigation in Antarctica and cooperation toward that end… shall continue
  Art. II
- Scientific observations and results from Antarctica shall be exchanged and made freely available
  Art. III

FAIR principles

Findable

Accessible

Interoperable

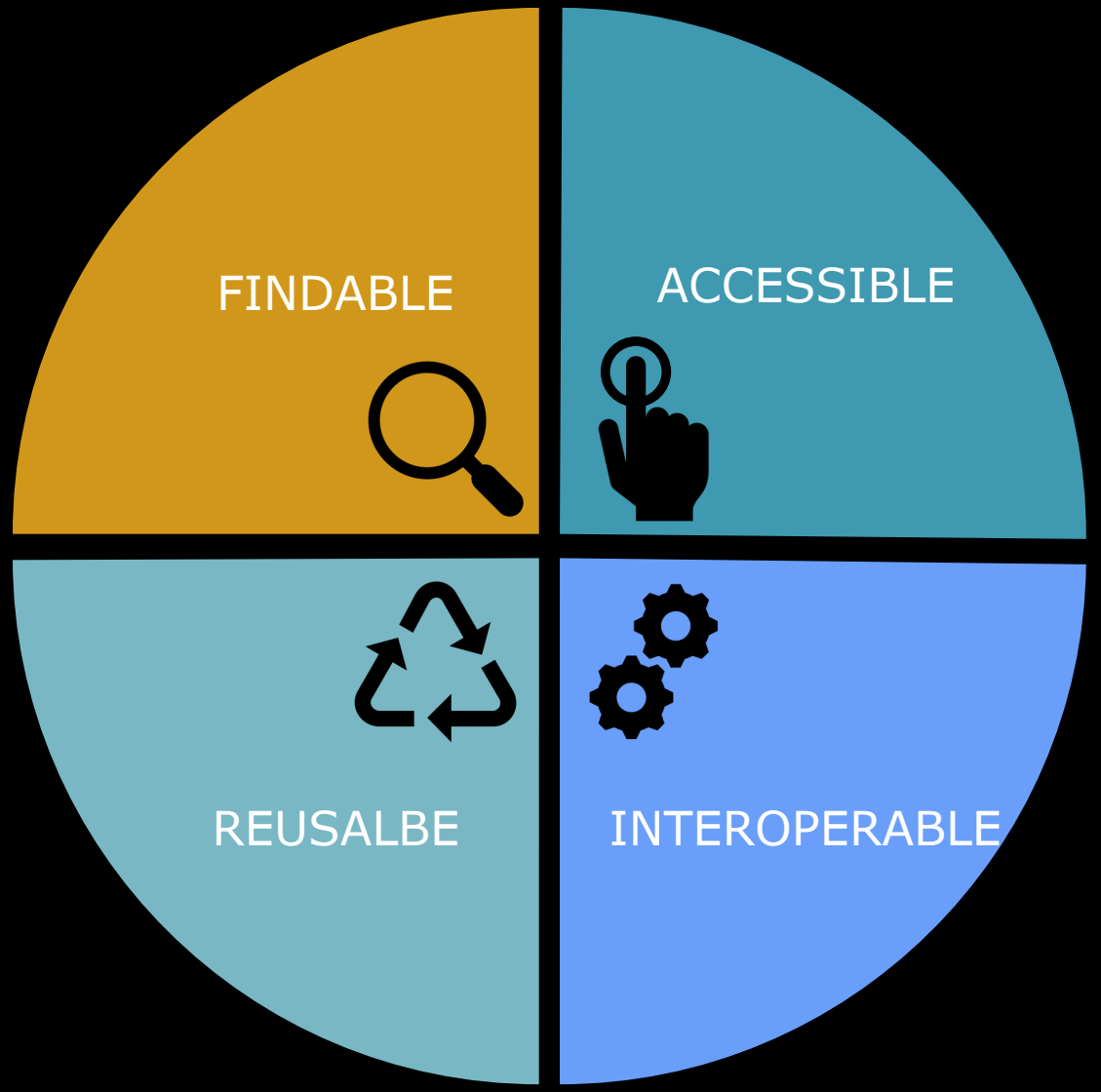
Reusable

- Metadata
  - Data about data
- Data
- Infrastructures

Machine readable

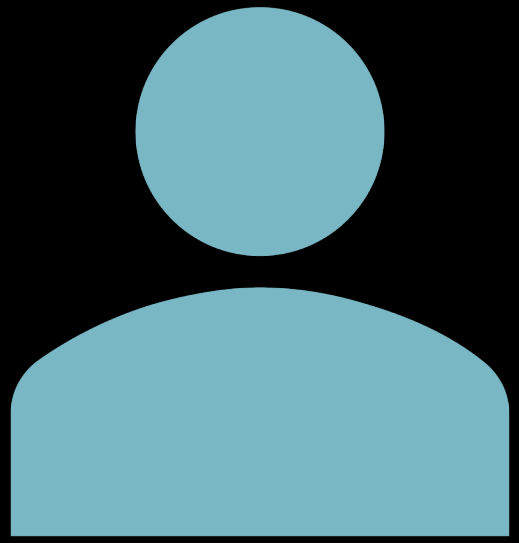
FAIR Principles

- Metadata
  - Data about data
- Data
- Infrastructures



FAIR is NOT

- not a standard
- not equal to ‘Open’ or ‘Free’
- no guarantee for successful (open) science



Fair Principles

MAKE YOUR DATA FAIR

Make a plan for the data before you start a project!

Collect detailed descriptive information (= metadata) throughout

Use standards and formats common to your discipline

Store the data in a trusted & sustainable repository or data center

See to that the data gets persistent identifiers (DOIs)

MAKE YOUR DATA FAIR

- Apply a suitable usage license
- Provide end users with information on “intended use”
- Make the data “as open as possible, as closed as necessary”
- Ensure that metadata remain available even if the data cannot be accessed any more
- FAIRness needs to be applied where it makes sense

Are your data FAIR?

- • ANDS FAIR self-assessment tool

https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/

- OzNome 5 Star Data Rating tool
- http://oznome.csiro.au/5star/
- • FAIRDAT assessment tool (prototype)
- https://www.surveymonkey.com/r/fairdat
- How FAIR is your data?

https://forms.gle/eBagszpWKVz5NKpp7

as open as possible and as closed as necessary

- EU Horizon 2020 framework
- Stems from the FAIR principles

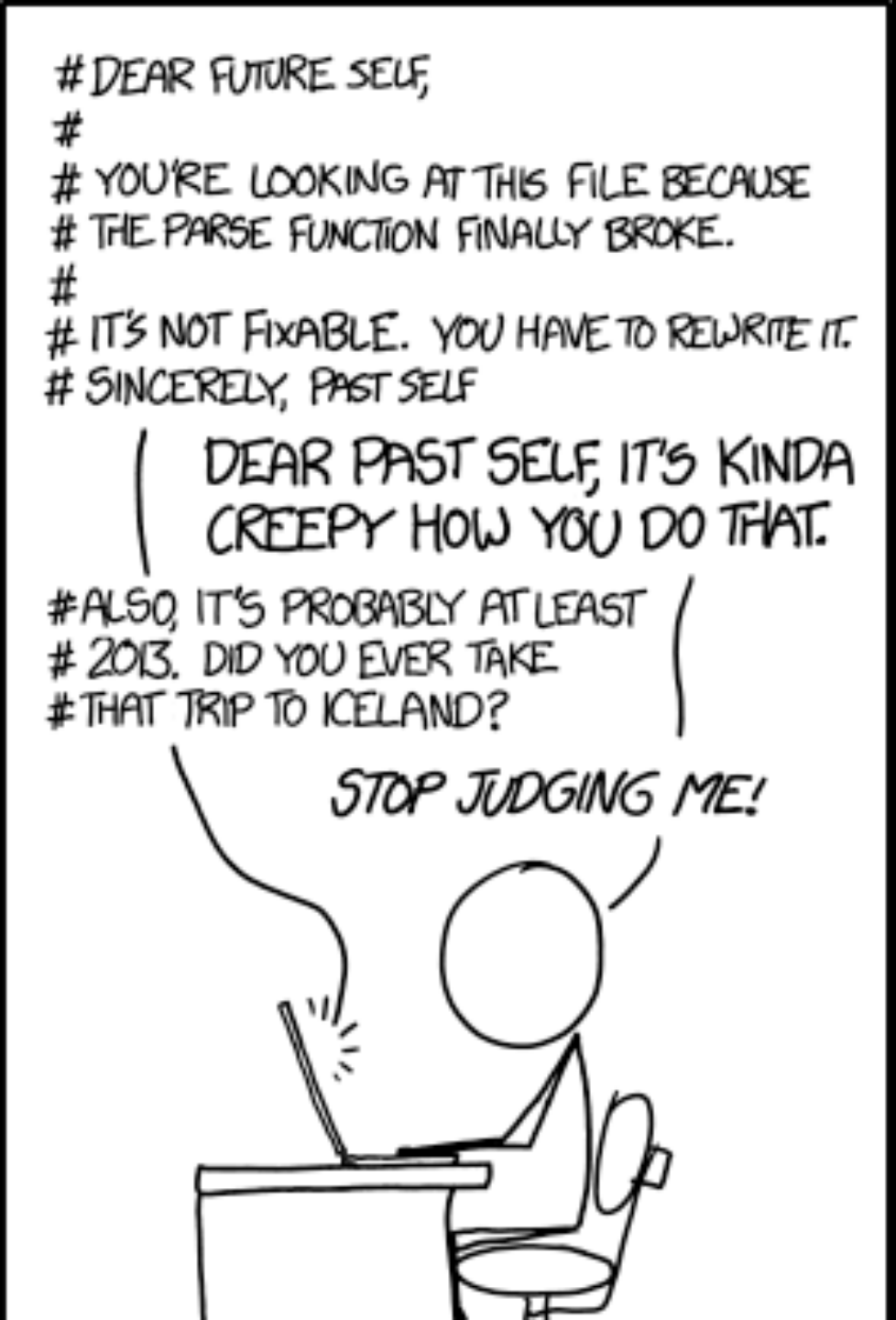
Open Reproducible Science

Transparency

Publicly available data and associated processing methods

Transparent communication of results

collaboration



Be the right kind of lazy

Start a data management plan

Do it right the first time

Think about future you

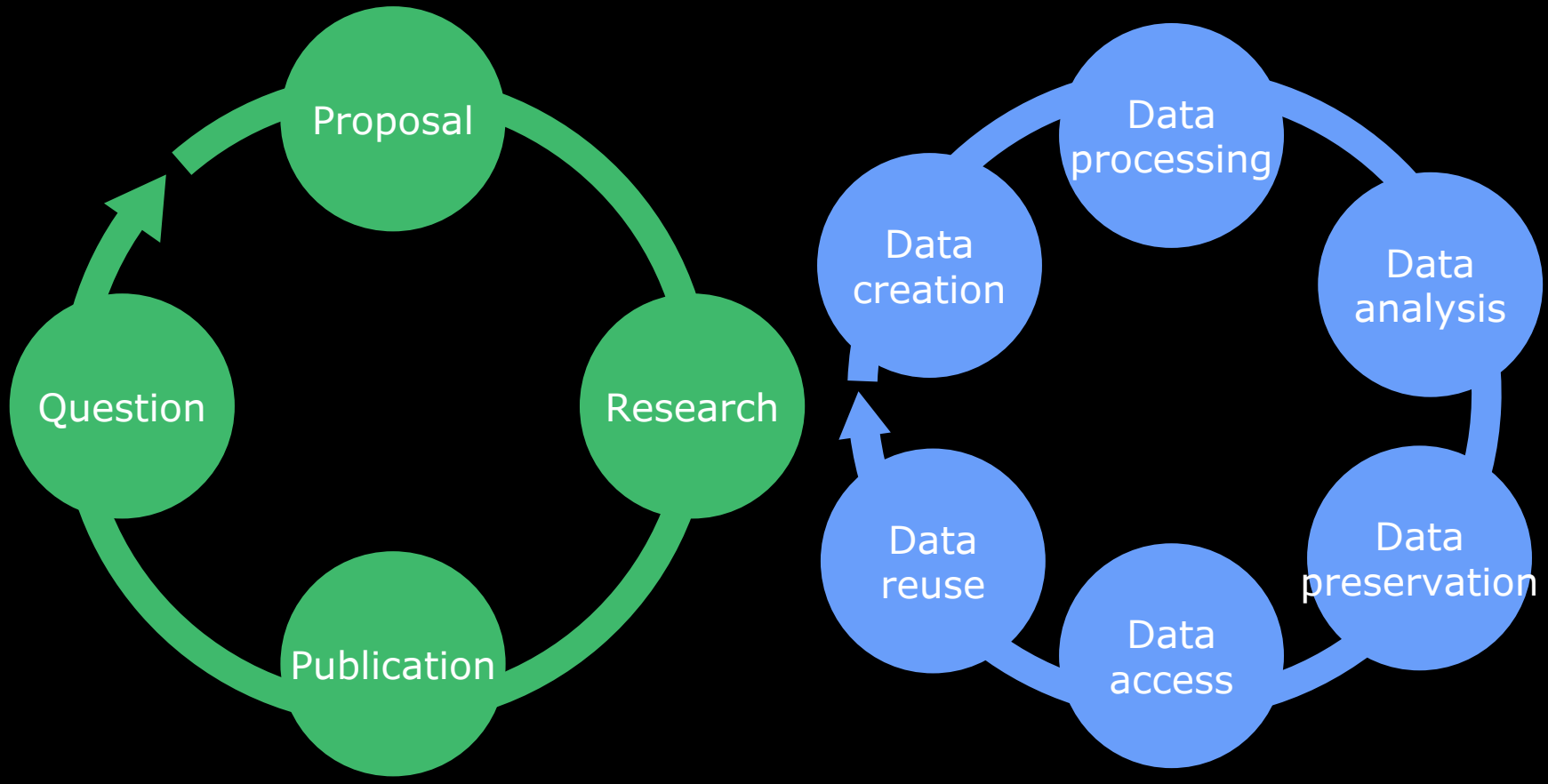
Describe your variables

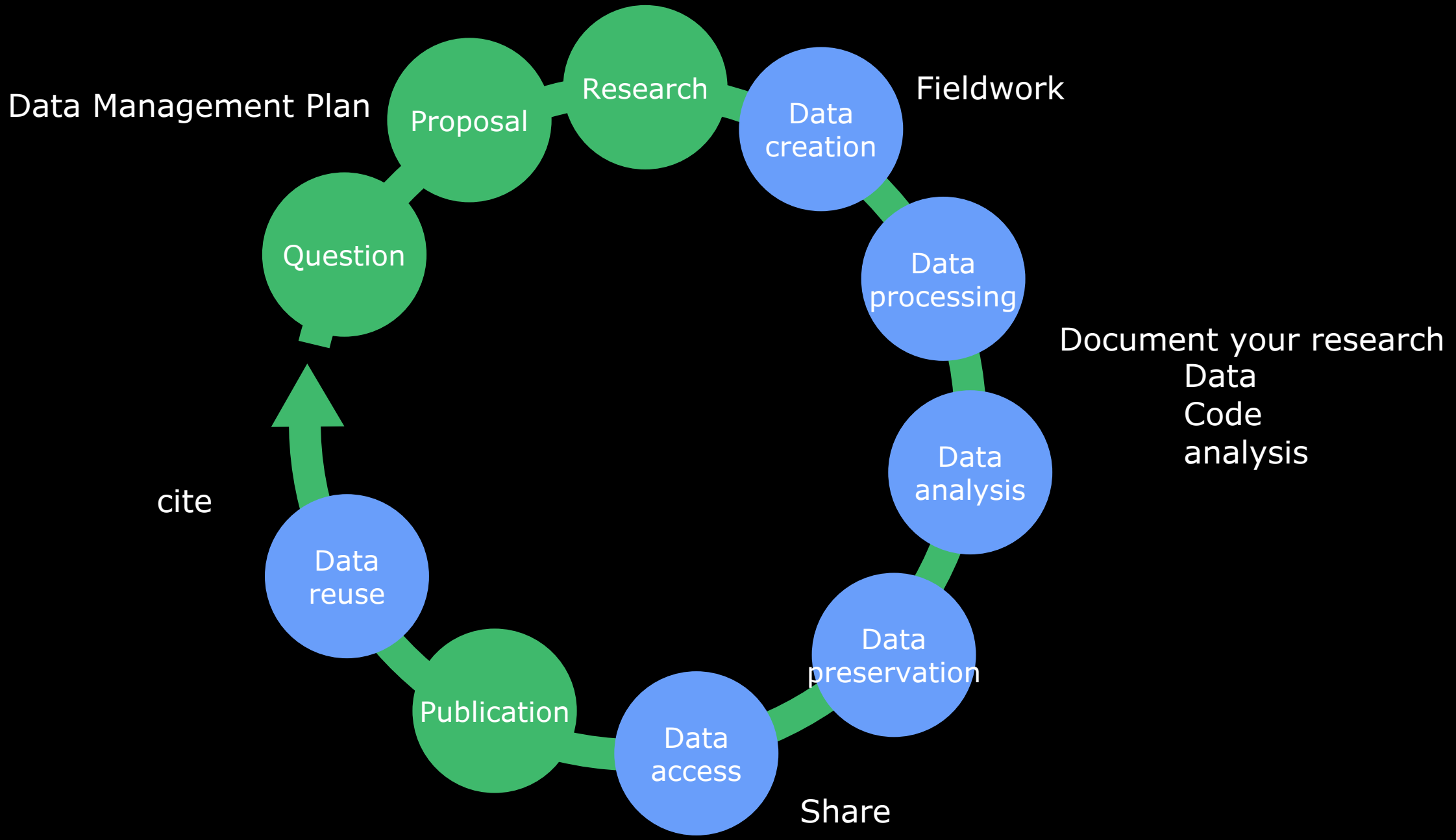
Make subsequent analysis easy

Roald Amundsen

- I may say that this is the greatest factor — the way in which the expedition is equipped — the way in which every difficulty is foreseen, and precautions taken for meeting or avoiding it. Victory awaits him who has everything in order — luck, people call it. Defeat is certain for him who has neglected to take the necessary precautions in time; this is called bad luck.
- Adventure is just bad planning

research and data life cycle




Data Management Plan
Proposal
Research
Data creation
Fieldwork
Question
Data processing
Document your research
Data
Code
analysis
Data analysis
cite
Data reuse
Data preservation
Publication
Data access
Share

Data management plan

- Good practices
- May be required by funders
- Various tools
- Living document

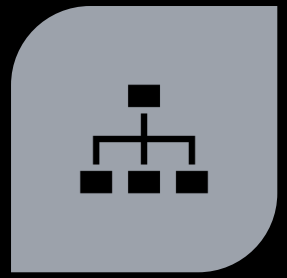
Data management plan

WHAT DATA WILL BE COLLECTED, AND HOW.

HOW DATA WILL BE DOCUMENTED.

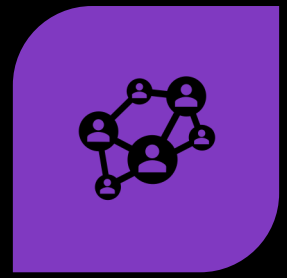
HOW ANY ETHICAL AND LEGAL ISSUES WILL BE DEALT WITH.

HOW DATA WILL BE STORED AND BACKED UP DURING RESEARCH.



WHO WILL BE RESPONSIBLE FOR DATA MANAGEMENT

PLANS FOR PRESERVING DATA BEYOND THE PROJECT'S END.



ANY PLANS FOR SHARING OR PROVIDING ACCESS TO DATA.

Resources

- DMPOnline - https://dmponline.dcc.ac.uk
- DMPTool - https://dmptool.org
- DMPEditor - https://my.usgs.gov/dmpeditor/
- ezDMP - https://ezdmp.org

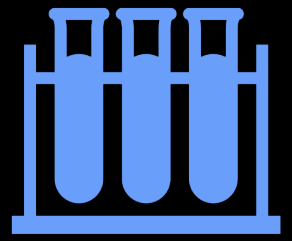
Prepare for the field

- Ask for help

- Make a list of what you need
  - Pack twice as much
  - Don’t assume it will be provided

- Start on time
  - Shipping can take a long time

WHAT DATA WILL BE
COLLECTED, AND HOW.



- Write out your sampling protocol
  - Test it

- Standardise your data

- Prepare templates
  - Paper is a valid medium for recording information

- Prepare labels
  - It saves time
  - Test them

Standardise your data

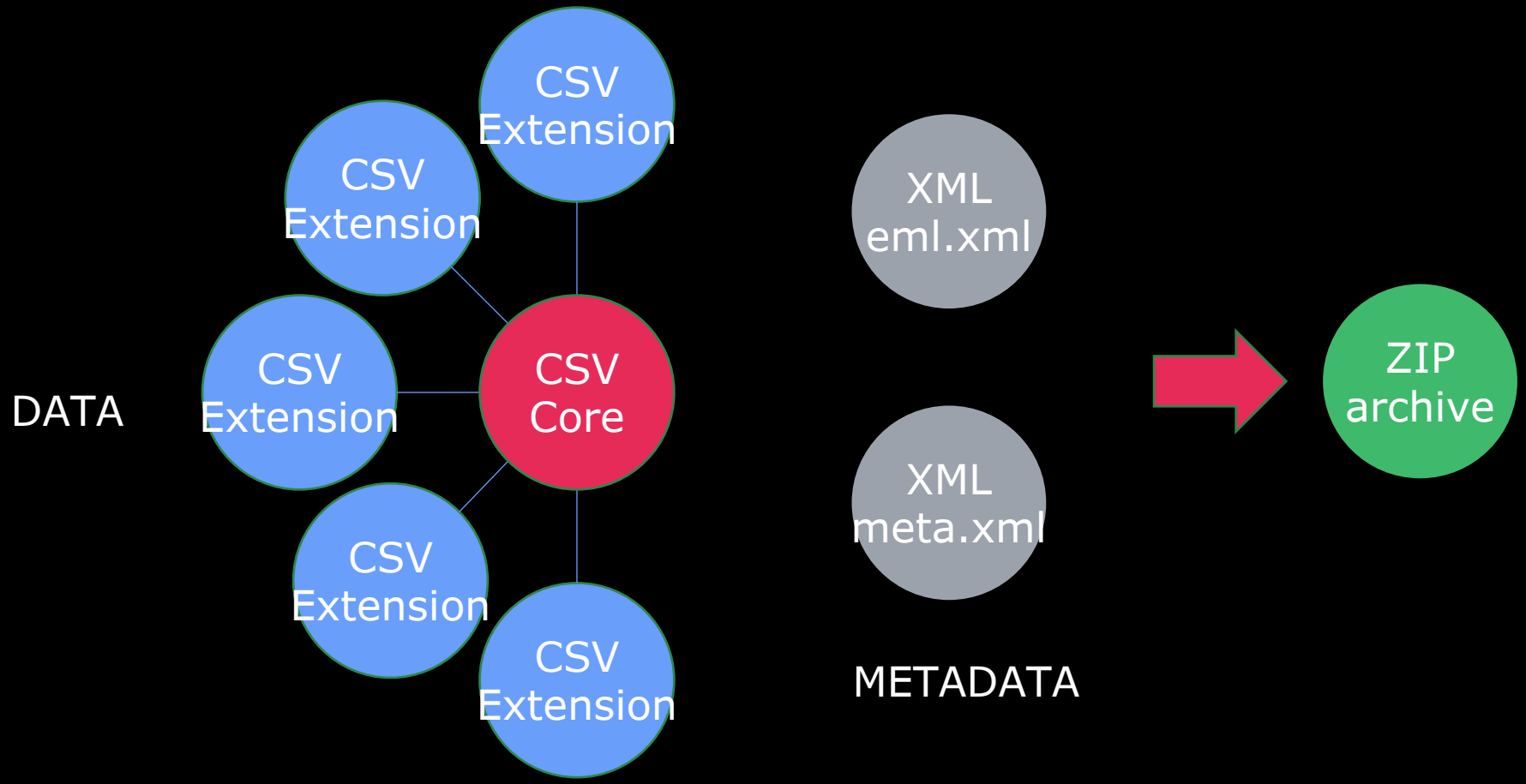
www.tdwg.org

- develops, ratifies and promotes standards and guidelines for the recording and exchange of data about organisms;
- acts as a forum for discussing all aspects of biodiversity information management through meetings, online discussions, and publications.

Darwin Core

- glossary of terms intended to **facilitate the sharing of information about biological diversity** by providing identifiers, labels, and definitions

Darwin Core standard serves to distribute biodiversity data



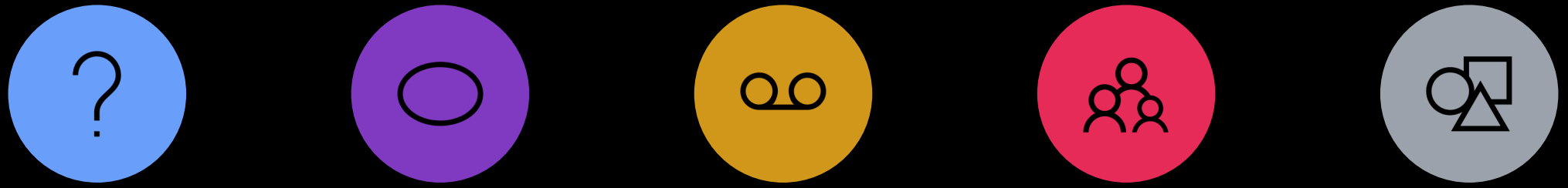
Darwin Core data types

METADATA
(NO CORE)

CHECKLIST

OCCURRENCE

EVENT

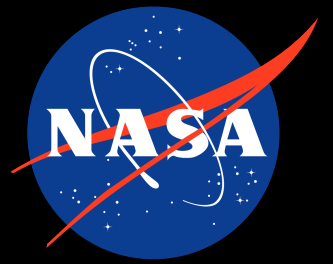


Find out more: Using a template to structure data: practical tips and tricks

Document your data

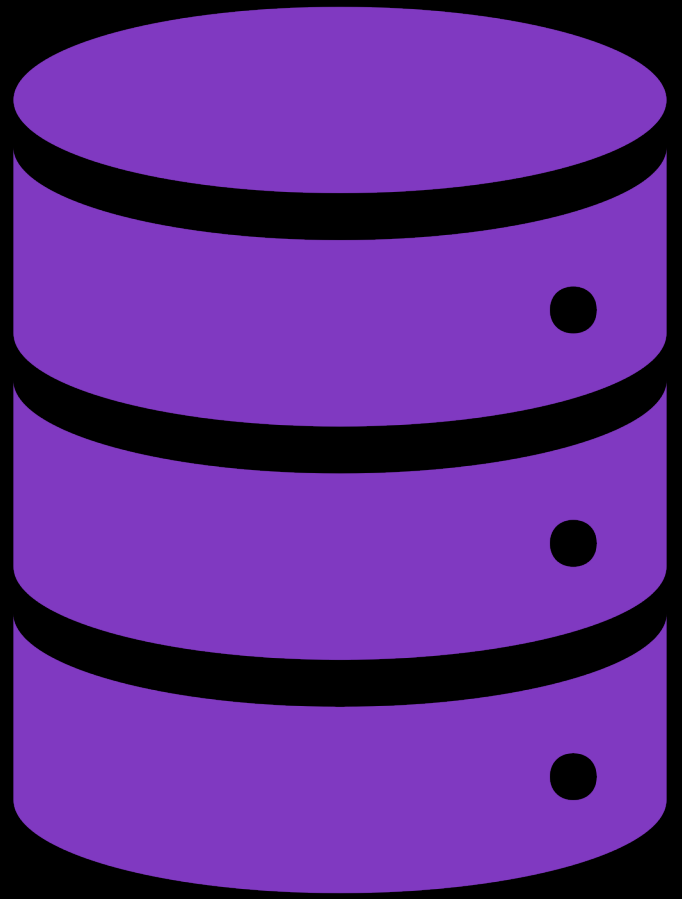
The Antarctic Master Directory

- Largest collection of Antarctic (meta)data
- > 9000 datasets
- > 25 countries



Some metadata formats

- DIF (AMD)
- ISO
- Schema.org
- EML Ecological Metadata Language (GBIF OBIS)



Ecological metadata Language

- Basic Metadata
- Geographic Coverage
- Taxonomic Coverage
- Temporal Coverage
- Keywords
- Associated Parties
- Project Data
- Sampling Methods
- Citations
- Collection Data

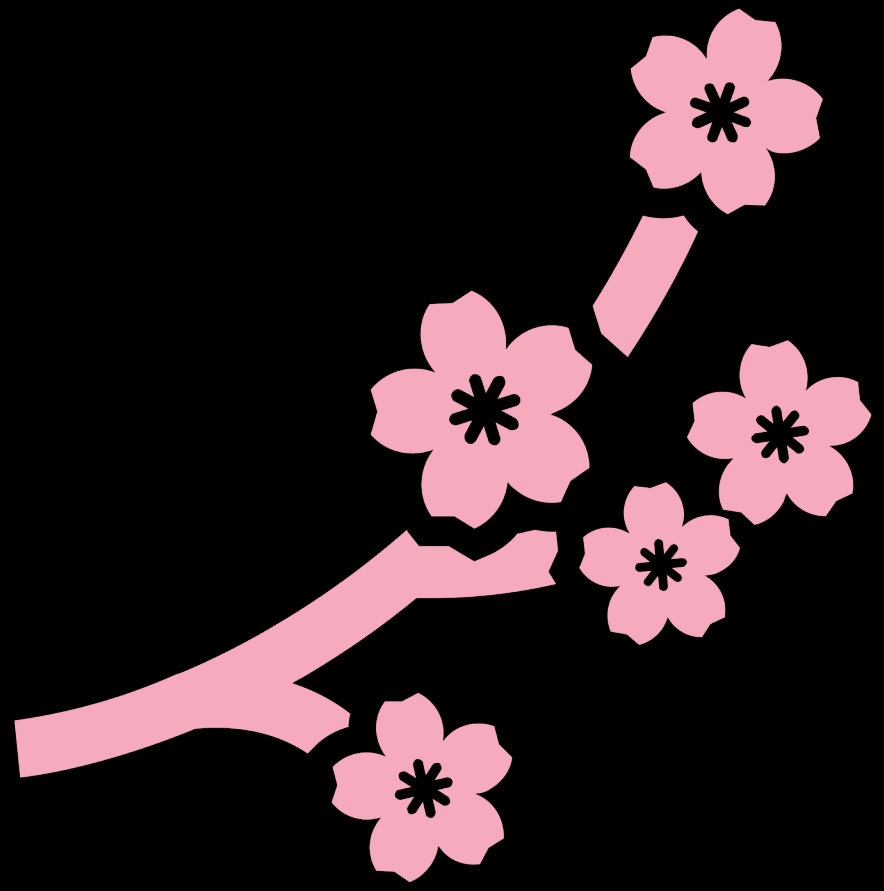
Metadata

- Materials and methods
- Abstract

HOW ANY ETHICAL AND LEGAL ISSUES WILL BE DEALT WITH.

Nagoya Protocol

not valid in international waters or Antarctica.



Madrid Protocol
Antarctic-Environmental Protocol

- impact assessment on the environment
- conservation of fauna and flora
- waste disposal and waste management
- prevention of marine pollution
- establishment and management of protected areas

Permits

- Check the permits you need
- Your own county
- The country you are travelling with
- The area you are traveling to

management of sensitive data

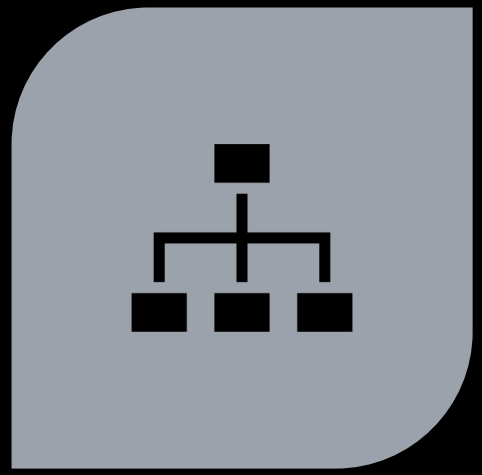
- integral to ethical data management.
- Wherever possible, environmental information should be freely available to all.
- can sometimes result in environmental harm.
- any restrictions should be assessed and reviewed rigorously.
- Documentation of reason(s) for the categorization is recorded as metadata that remains with the record.

Data management plan

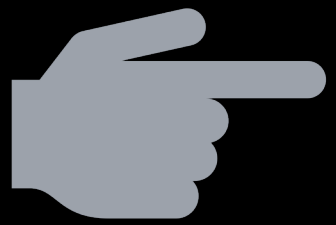
HOW DATA WILL BE STORED AND BACKED UP DURING RESEARCH.

In the field

- Notebooks
- Paper
- Pictures
- Hard Drive
- Cloud



WHO WILL BE RESPONSIBLE
FOR DATA MANAGEMENT



YOU!!

- (Natural History) institute
- University
- National Antarctic Data Centres
- Thematic data centres

PLANS FOR PRESERVING DATA
BEYOND THE PROJECT'S END.

Find a repository

National Antarctic Data Centres (SCADM)

www.biodiversity.aq

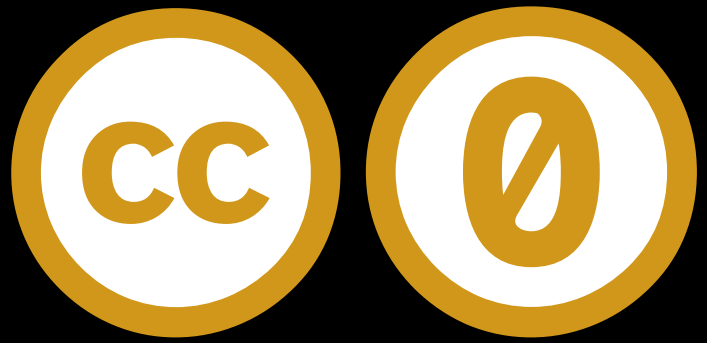
ZENODO

ANY PLANS FOR SHARING OR
PROVIDING ACCESS TO DATA.

Choose a license

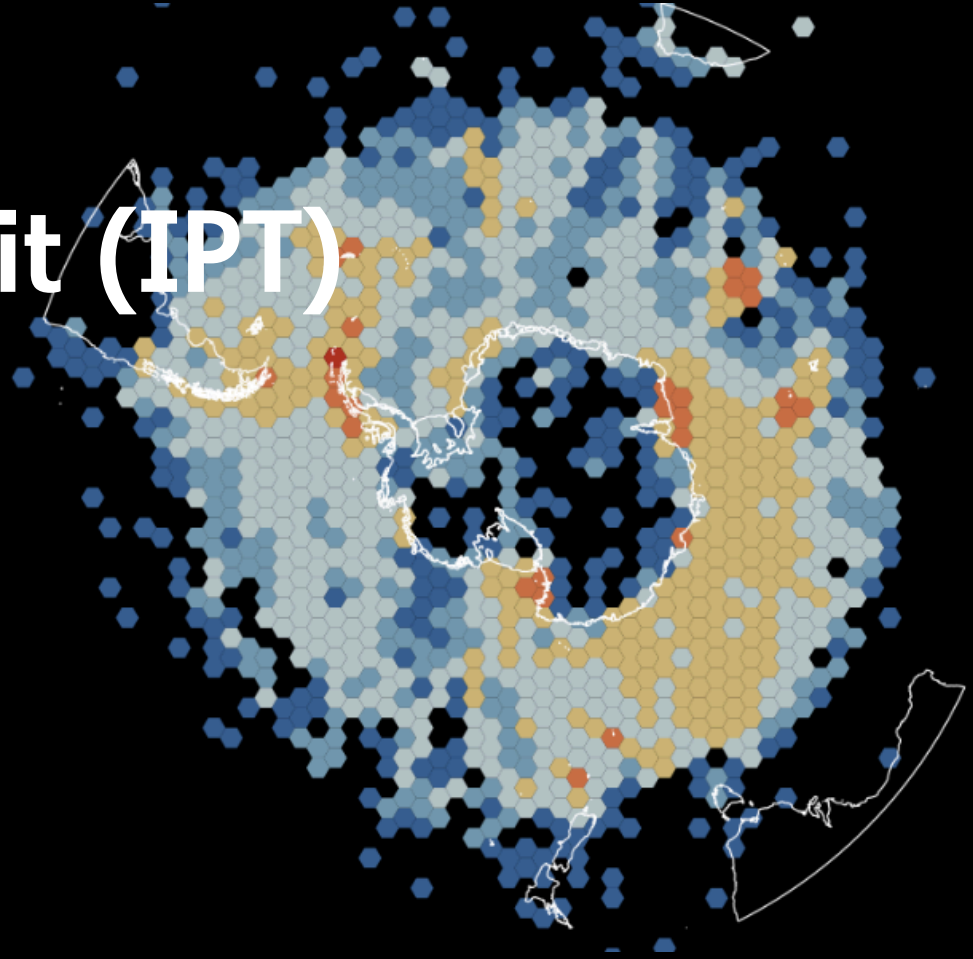
Use Creative Commons Zero

- Creative Commons Zero is the most appropriate license for scientific (biodiversity) data
  - reduce any legal and technical impediments, be they intentional and unintentional, to the reuse of data
  - does not exempt those who reuse the data from following community norms for scholarly communication
- If you must use CC-by or CC-by-NC



Publish your data

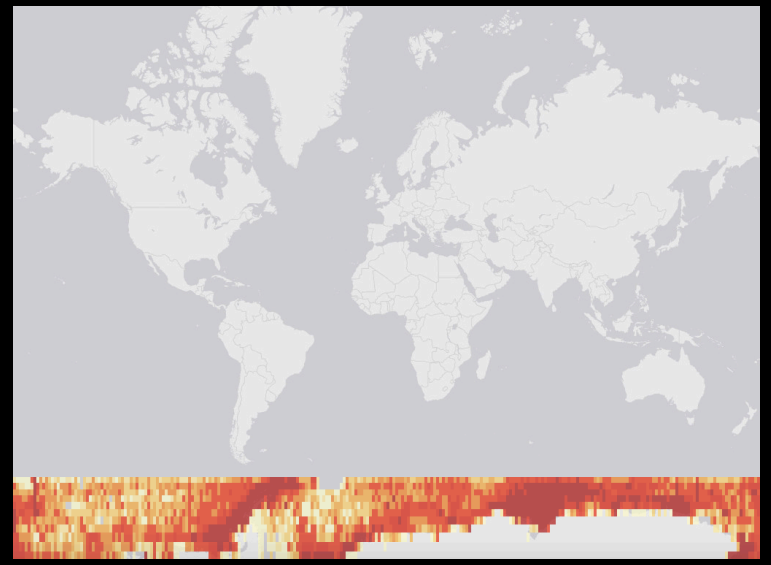
Integrated Publishing Toolkit (IPT)



- Easiest and most interoperable way to publish data
- IPT.biodiversity.aq
- Linked to largest biodiversity information Facilities
- data.biodiversity.aq

OBIS

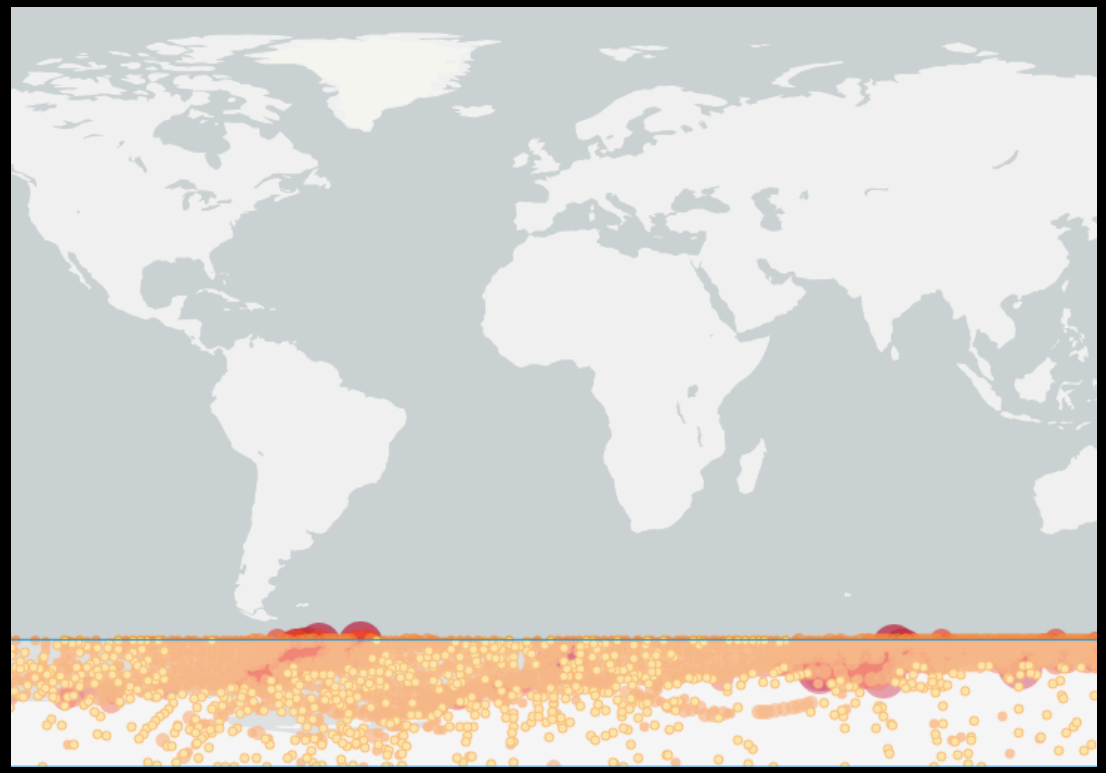
OBIS



Ocean Biogeographic Information System

Global Biodiversity Information Facility

- -60°-90°
- 1,784,919 records



Findability and Accessibility

Registered datasets get a DOI, are findable through the websites and their APIs

Open data promotes analytical re use

- Examples see EG-ABI
- https://www.scar.org/science/egabi/
- Reproducible workflows for reproducible science
- https://ropensci.org

Closing thought: Cite the DOI

Cite the data you use

Cite the code you use

SCAR data Laundry Slack Channel

https://join.slack.com/t/scar-data-laundry/shared_invite/zt-e7b42oi8-N0nP8k4TxTfl36r873R0~g

THANK YOU!

METADATA (NO CORE)

CHECKLIST

OCCURRENCE

EVENT

Darwin Core data types

Findable

- Does the dataset have any identifiers assigned?
- Is the dataset identifier included in all metadata records/files describing the data?
- How is the data described with metadata?
- What type of repository or registry is the metadata record in?

DOI

mostly

eml.xml

domain
+ general

Accessible

- **How accessible is the data?**
- **Is the data available online without requiring specialised protocols or tools once access has been approved?**
- **Will the metadata record be available even if the data is no longer available?**

Public

API + Online

Not necessarily

Interoperable

- **What (file) format(s) is the data available in?**
- **What best describes the types of vocabularies/ontologies/tagging schemas used to define the data elements?**
- **How is the metadata linked to other data and metadata (to enhance context and clearly indicate relationships)?**

Structured, open, machine readable

suggested

Reusable

- **Which of the following best describes the license/usage rights attached to the data?**
- **How much provenance information has been captured to facilitate data reuse?**