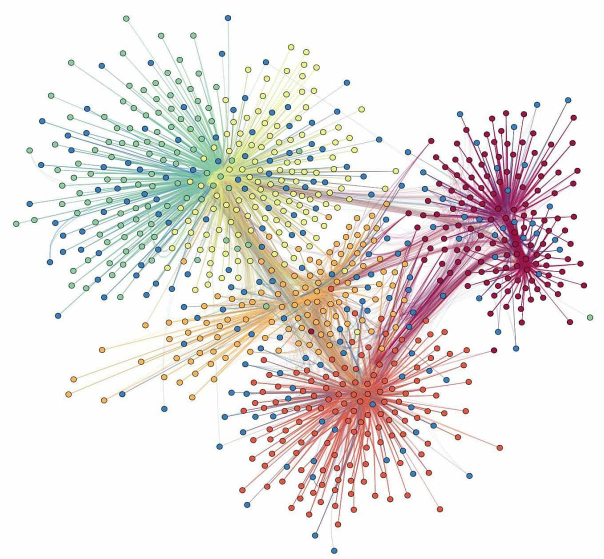
Big Data Ecology

Fundamentals of reproducible and collaborative ecological research



Christian König

Ecology and Macroecology Lab
Institute for Biochemistry and Biology
University of Potsdam

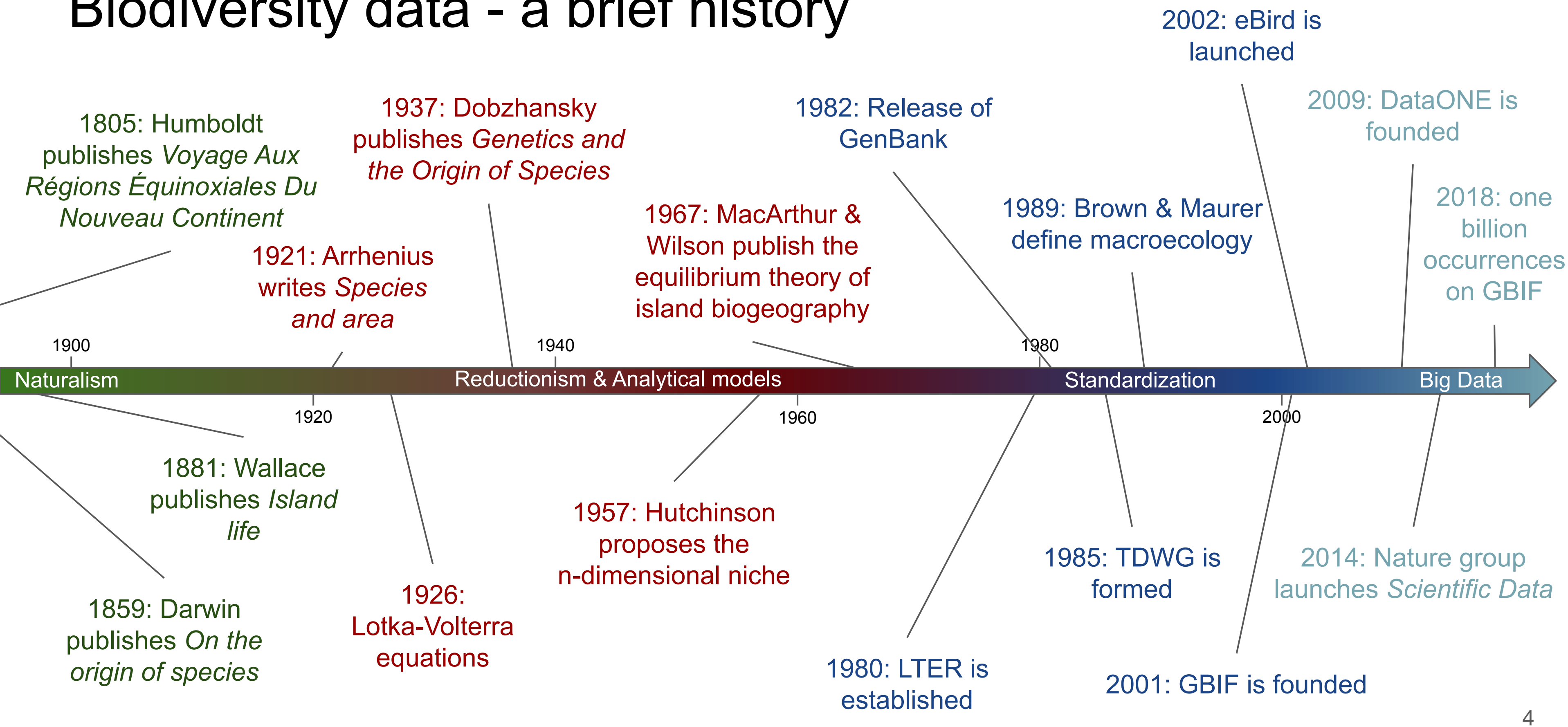
Aims

- Getting an overview of the principles of modern ecological research and the available infrastructure to support it
- Learn about tools and techniques that enable the collation, integration and analysis of large ecological datasets
- Gain hands-on experience with different aspects of the ecological data life cycle through a practical project
- Be able to transfer this knowledge to your own PhD research project

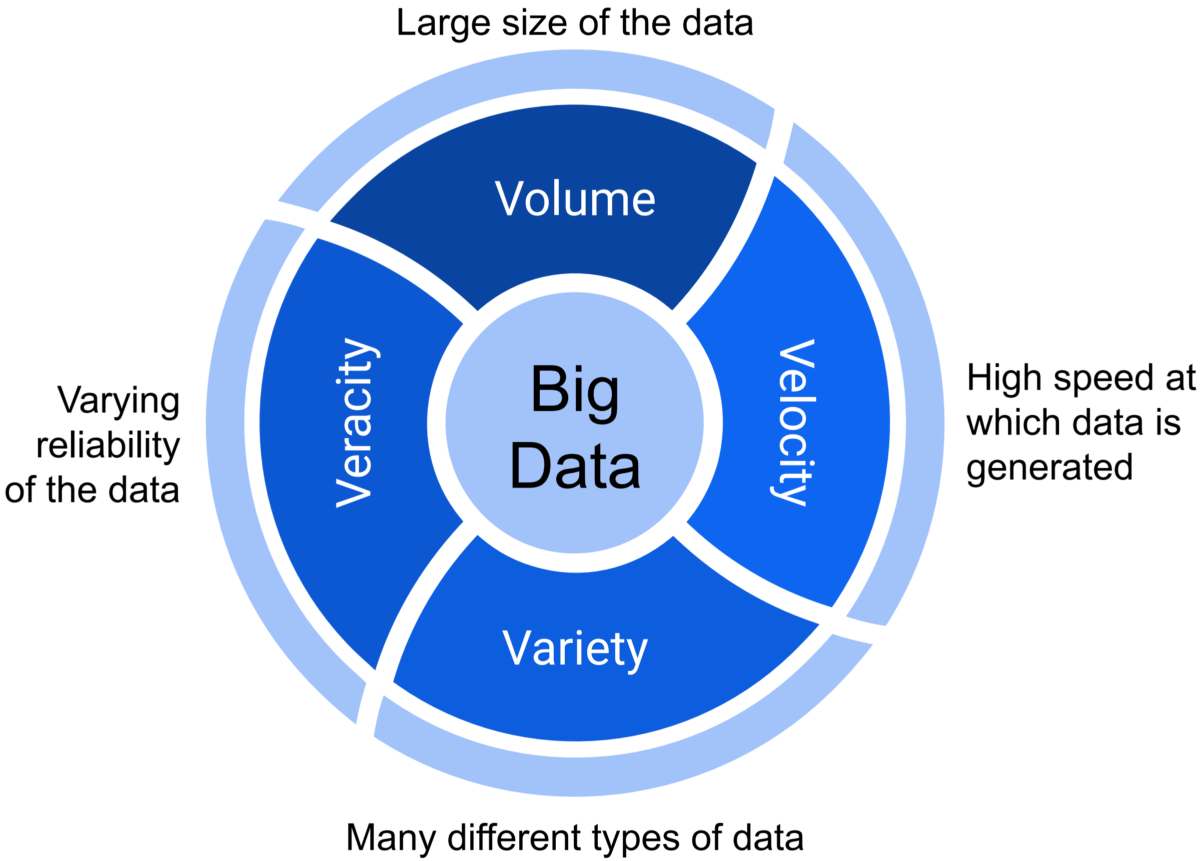
Schedule

	Day 1	Day 2	Day 3
Unit 1 9:00-10:30	Welcome & Introductory round Introductory Lecture	Project [Collect, Assure] Lecture: Spatial data	Project [Integrate] Discussion: Analysis and Visualizaiton
Unit 2 11:00-12:30	Lecture: Version Control Practical: Git basics Project [Plan]	Practical: Working with spatial data Project [Discover]	Project [Analyze, Publish] Wrap up & final steps, Concluding discussion
Unit 3 13:30-15:00	Lecture: Data bases Practical: SQL, dplyr, etc. Project [Collect, Assure]	Lecture: Computional performance Practical: Performance & Parallel processing	

Biodiversity data - a brief history



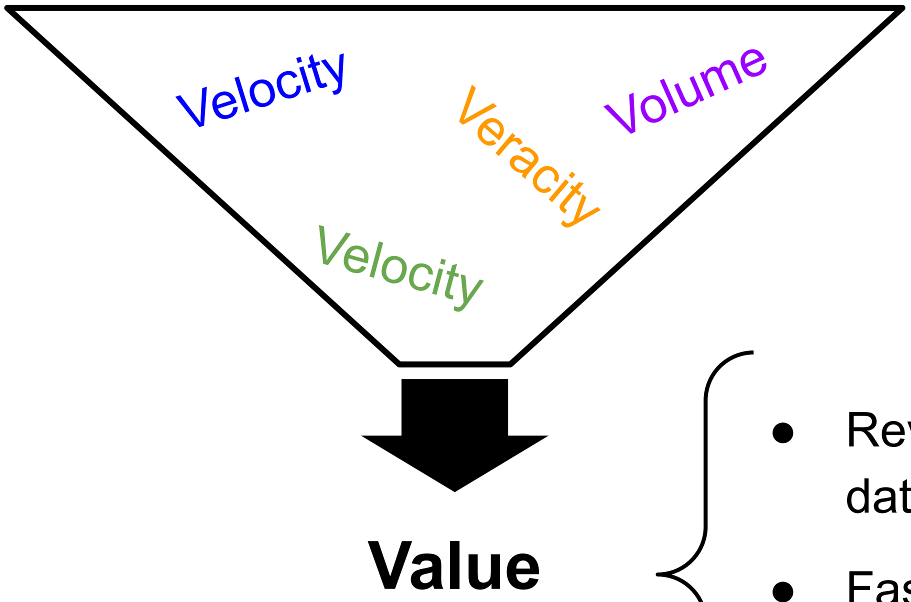
What is "Big Data"



"Even the smallest datasets can contribute key knowledge for large-scale problem solving"

Hampton et al. (2013)
10.1890/120103

The promise of “Big Data”



- Reveal patterns that are hidden in individual datasets
- Faster, more reliable inferences
- More specific and accurate predictions

Unlocking the potential of Big Data - The FAIR data principle

Findable, **A**ccessible, **I**nteroperable and **R**eusable

Why?

- Science is a collaborative, incremental process
- Increased research impact
- Acknowledgement of other people's work

SCIENTIFIC DATA

Amended: Addendum

OPEN

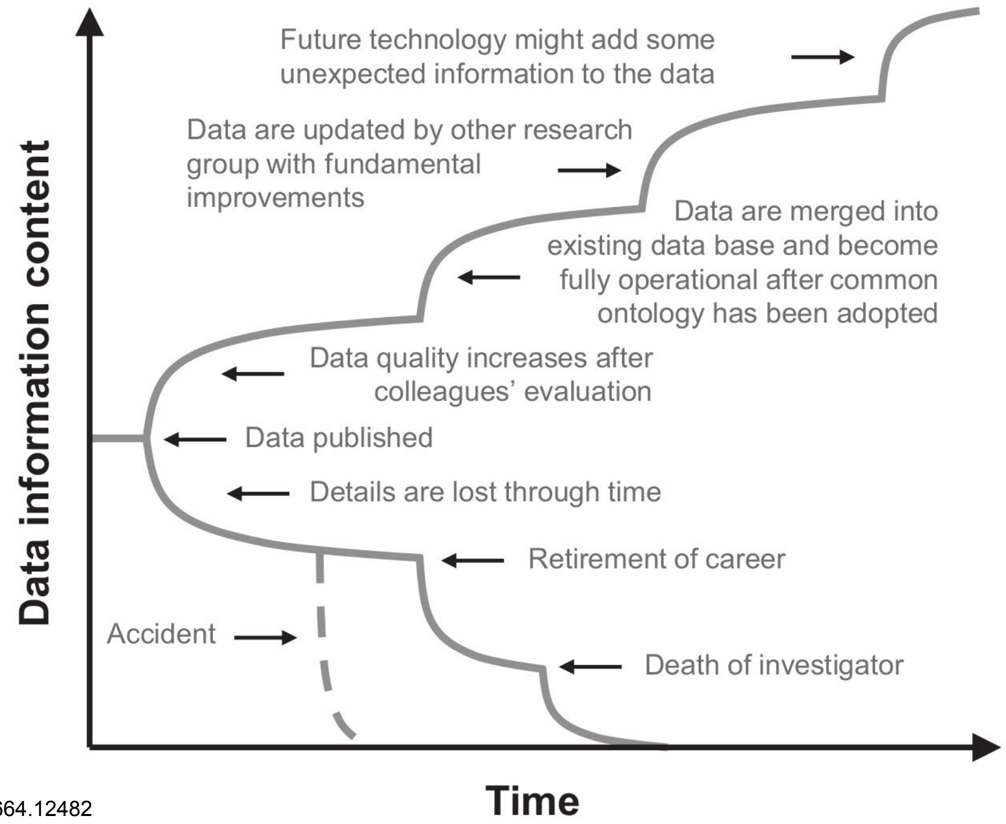
SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

Wilkinson et al. (2016) 10.1038/sdata.2016.18

The FAIR data principle



Source: Mouquet et al. (2015) 10.1111/1365-2664.12482

The FAIR data principle

Findable

F1: (meta)data are assigned a globally unique and eternally persistent identifier.

F2: data are described with rich metadata.

F3: (meta)data are registered or indexed in a searchable resource.

F4: metadata specify the data identifier.

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

The FAIR data principle

Accessible

A1: (meta)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.

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

The FAIR data principle

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.

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

The FAIR data principle

Reusible

R1: meta(data) have 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 their provenance.

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

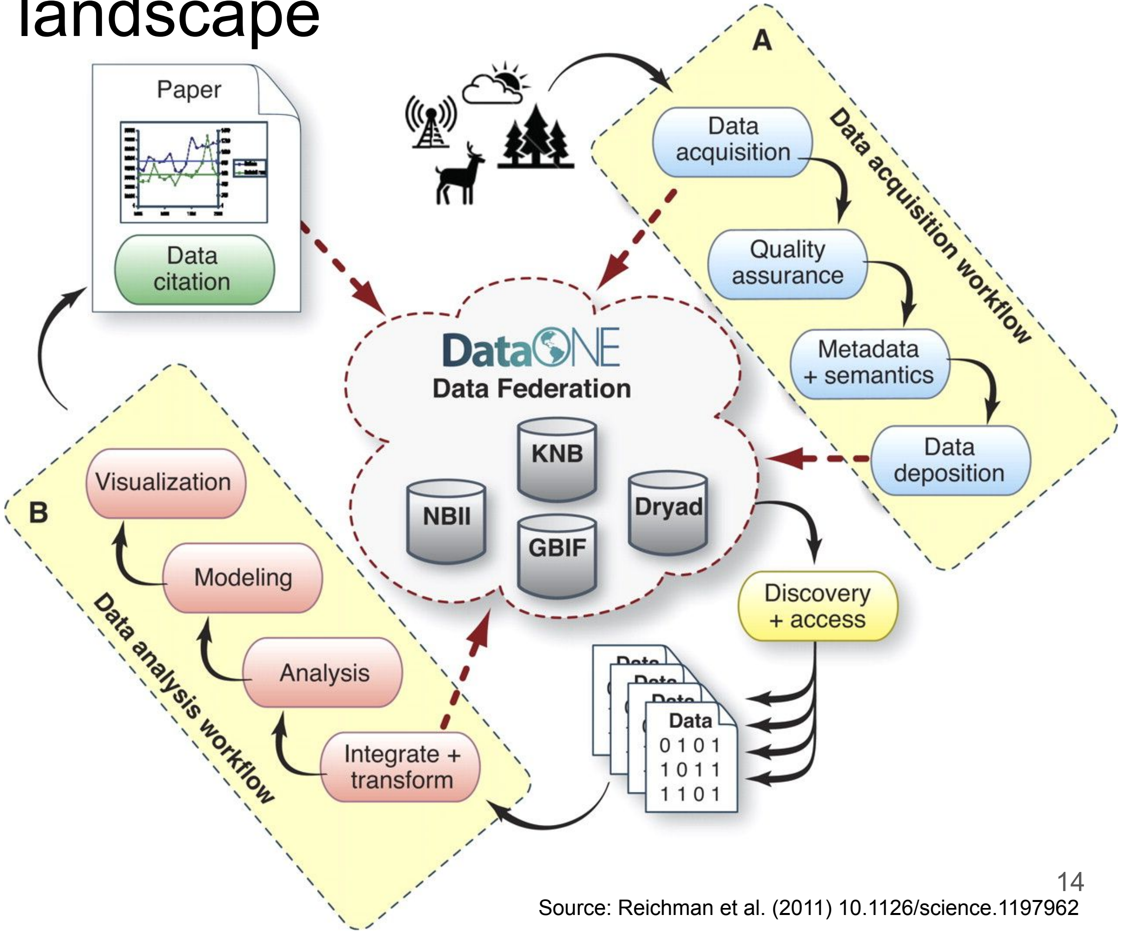
Source: https://www.go-fair.org/fair-principles/

What are key terms mentioned in the FAIR principles

A FAIR biodiversity data landscape

Three key components:

- Common standards and protocols
- Federated architecture
- Integrated data lifecycle



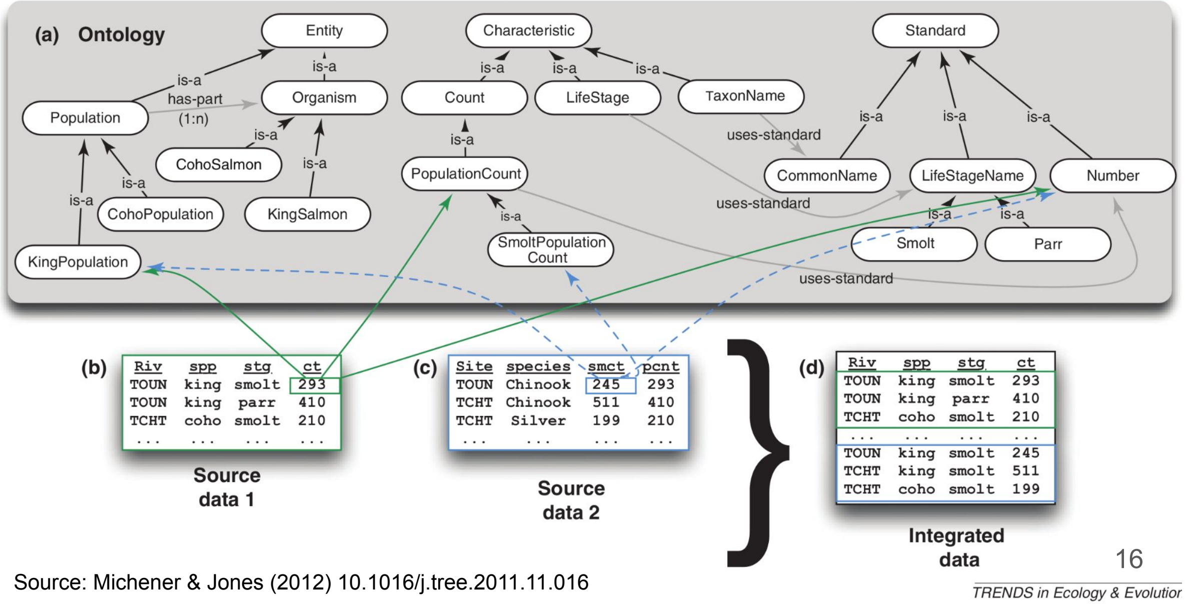
Source: Reichman et al. (2011) 10.1126/science.1197962

A FAIR biodiversity data landscape - Standards & Protocols

- Vocabularies:
    - Controlled terminology to describe a given concept
    - Examples: Essential Biodiversity Variables (conceptual), GBIF backbone (taxonomy), TOP thesaurus (traits), LTER controlled vocabulary (general)
- Identifiers:
    - Unique and persistent *name* for an object
    - Examples: DOI, ORCID, LSID
- Data formats:
    - Standardized (but often flexible!) ways to ecological represent data
    - Examples: Darwin Core Archive, Ecological Metadata language, Access to Biological Collection Data

A FAIR biodiversity data landscape - Standards & Protocols

- Ontologies
  - Formal representation of the relationships between and properties of different entities
  - Often domain-specific
  - Examples: OBO, ENVO



Riv	spp	stg	ct
TOUN	king	smolt	293
TOUN	king	parr	410
TCHT	coho	smolt	210
...	...	...	...

Site	species	smct	pcnt
TOUN	Chinook	245	293
TCHT	Chinook	511	410
TCHT	Silver	199	210
...	...	...	...

Riv	spp	stg	ct
TOUN	king	smolt	293
TOUN	king	parr	410
TCHT	coho	smolt	210
...	...	...	...
TOUN	king	smolt	245
TCHT	king	smolt	511
TCHT	coho	smolt	199

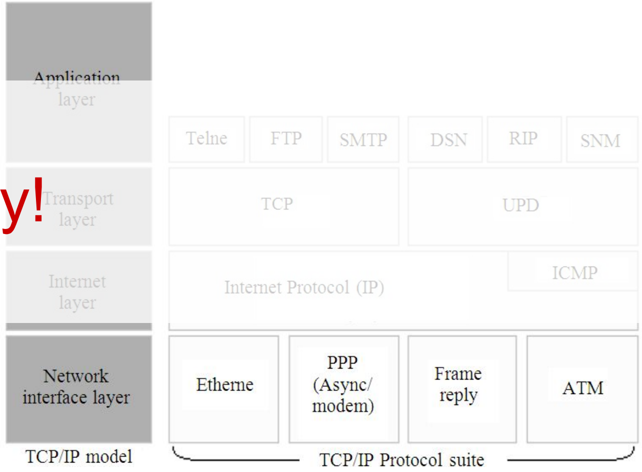
Source: Michener & Jones (2012) 10.1016/j.tree.2011.11.016

A FAIR biodiversity data landscape - Standards & Protocols

International System of Units (SI)

A success story!

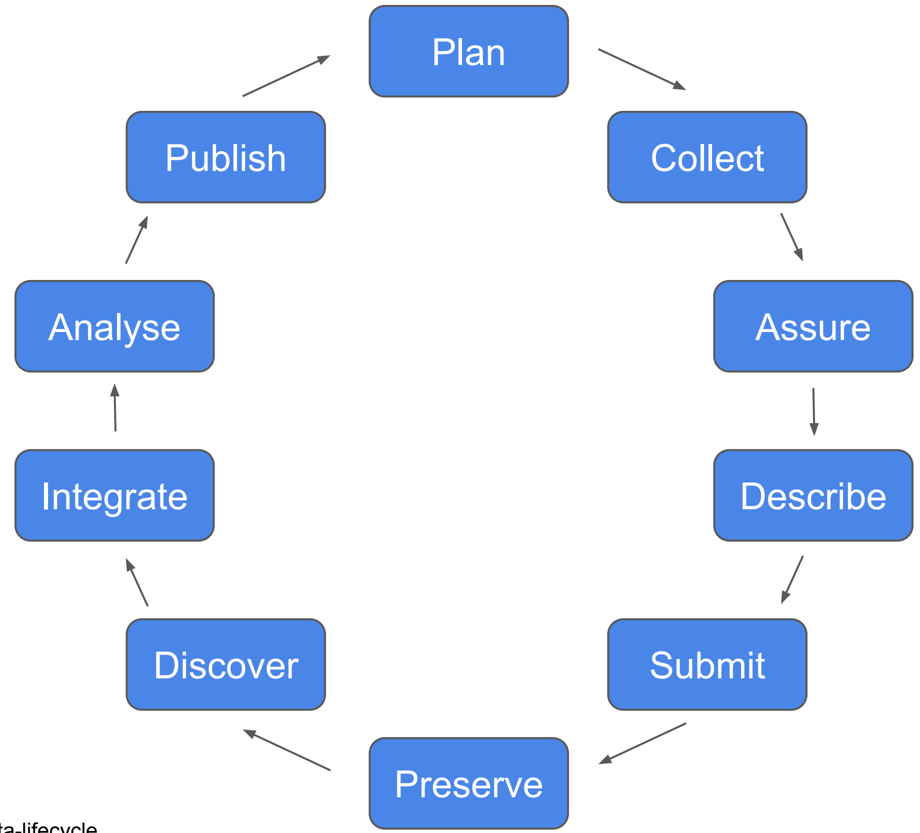
The Internet



A FAIR biodiversity data landscape - Federation

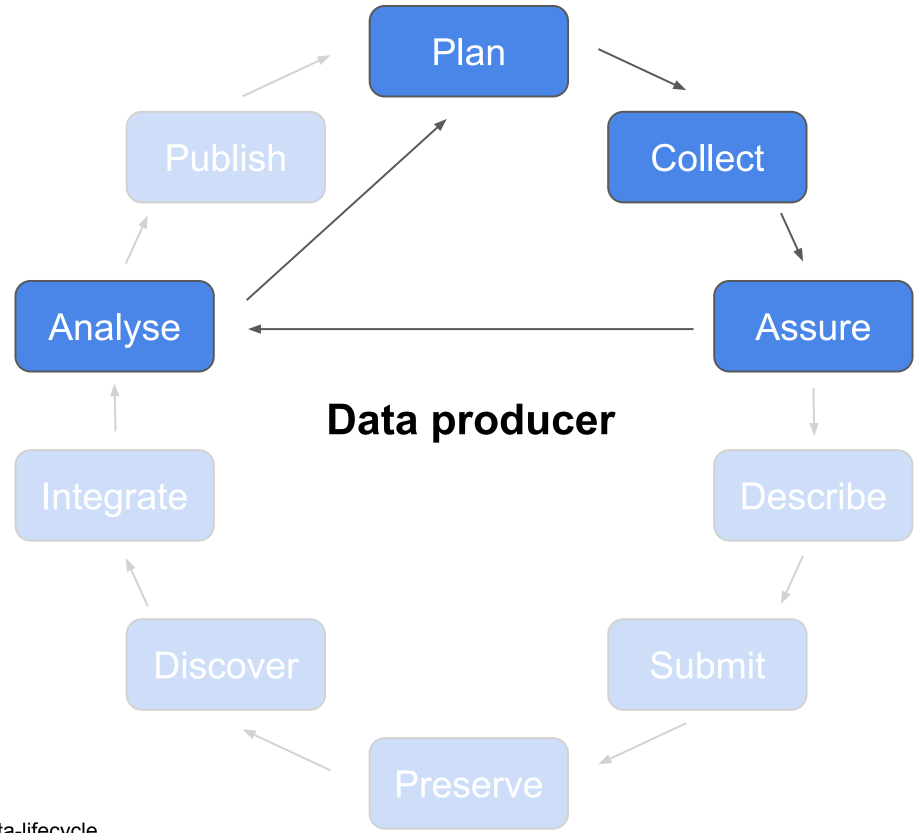
- Distributed management of data by member nodes with proven domain-expertise
- Integration of domain-specific repositories in a *network of databases*
- Search, discovery, access and use of data from different repositories via a common interface
- Advantages:
  - Distributed workload, computation and decision power
  - Data is managed by domain-experts
  - Easy attribution and provenance tracking
  - Higher engagement of the research community
- Examples: DataONE, GBIF

A FAIR biodiversity data landscape - The data life cycle



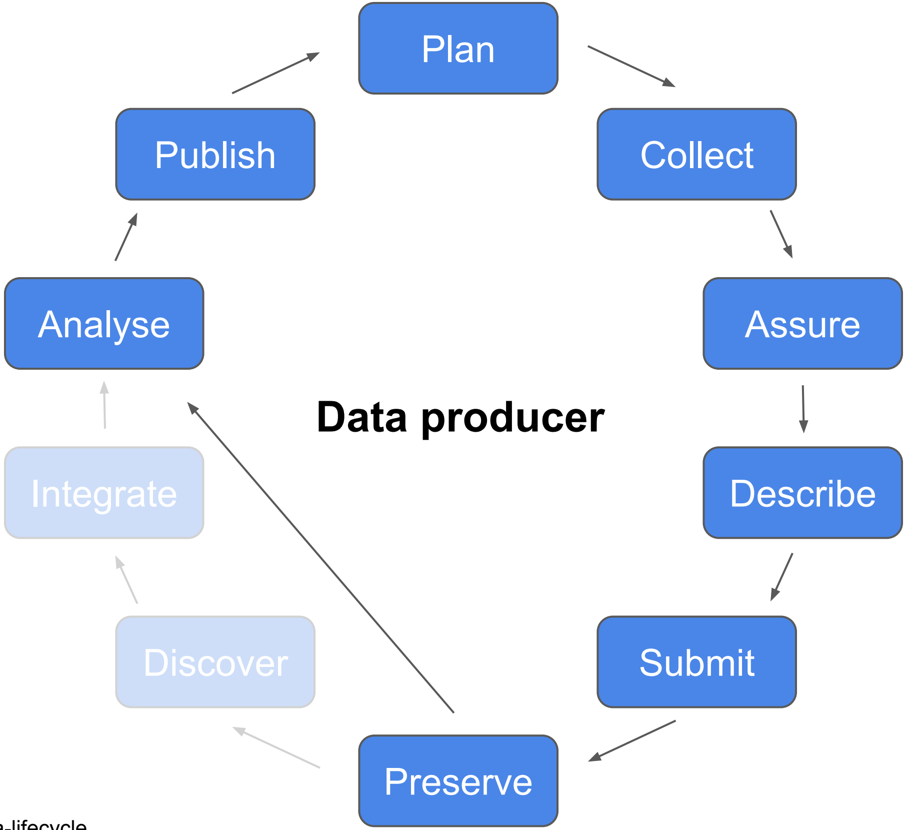
Source: https://www.gfbio.org/training/materials/data-lifecycle

A FAIR biodiversity data landscape - The data life cycle



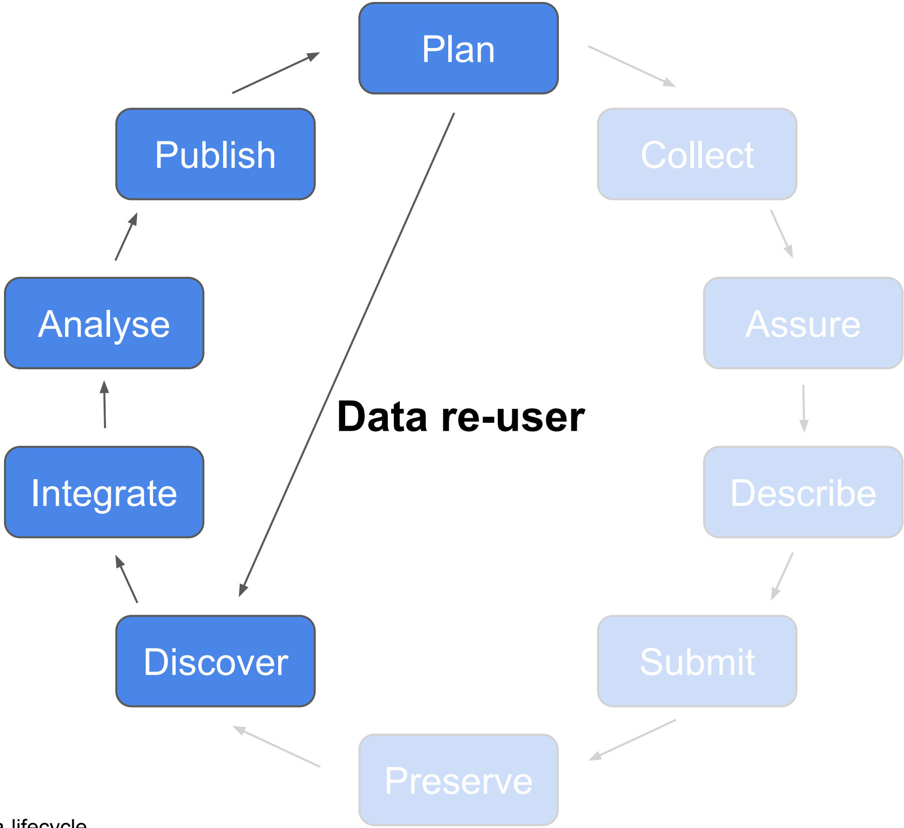
Source: https://www.gfbio.org/training/materials/data-lifecycle

A FAIR biodiversity data landscape - The data life cycle



Source: https://www.gfbio.org/training/materials/data-lifecycle

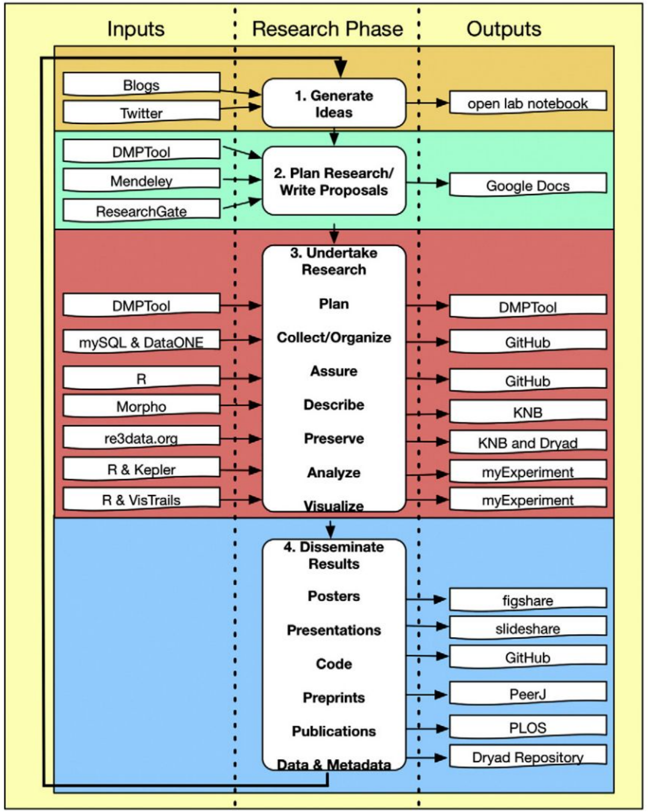
A FAIR biodiversity data landscape - The data life cycle



Source: https://www.gfbio.org/training/materials/data-lifecycle

A FAIR biodiversity data landscape - The data life cycle

There are many tools, resources and projects that can help you to structure your research according to FAIR principles!



Source: Michener (2015) 10.1016/j.ecoinf.2015.06.010

Further readings

Farley, S. S. et al. 2018. Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. - BioScience 68: 563–576.

Hampton, S. E. et al. 2013. Big data and the future of ecology. - Front. Ecol. Evol. 11: 156–162.

Hampton, S. E. et al. 2015. The Tao of open science for ecology. - Ecosphere 6: 120.

Hardisty, A. R. et al. 2019. The Bari Manifesto: An interoperability framework for essential biodiversity variables. - Ecological Informatics 49: 22–31.

Michener, W. K. 2015. Ten Simple Rules for Creating a Good Data Management Plan. - PLOS Computational Biology 11: 1004525.

Michener, W. K. and Jones, M. B. 2012. Ecoinformatics: Supporting ecology as a data-intensive science. - TREE 27: 85–93.

Michener, W. et al. 2011. DataONE: Data Observation Network for Earth - Preserving Data and Enabling Innovation in the Biological and Environmental Sciences. - D-Lib Magazine in press.

Reichman, O. J. et al. 2011. Challenges and opportunities of open data in ecology. - Science 331: 703–5.

Research project - Overview

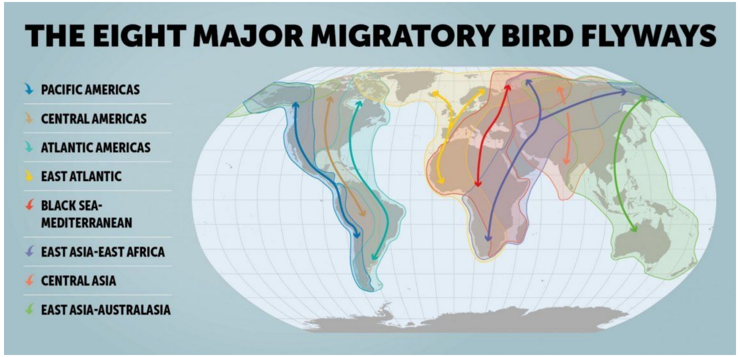
Aim: (1) Explore the imprint of seasonal bird migration in digitally available occurrence records

(2) Contrast patterns among congeneric species with different migration behaviour

Study taxon: Harriers (*Circus* Lacepede 1799)

Study region: Sweden

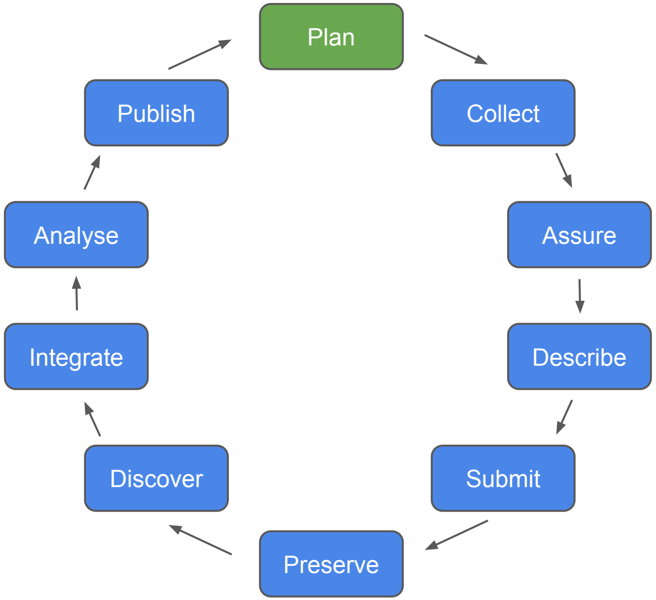
Circus cyaneus in flight. Source: wikimedia commons



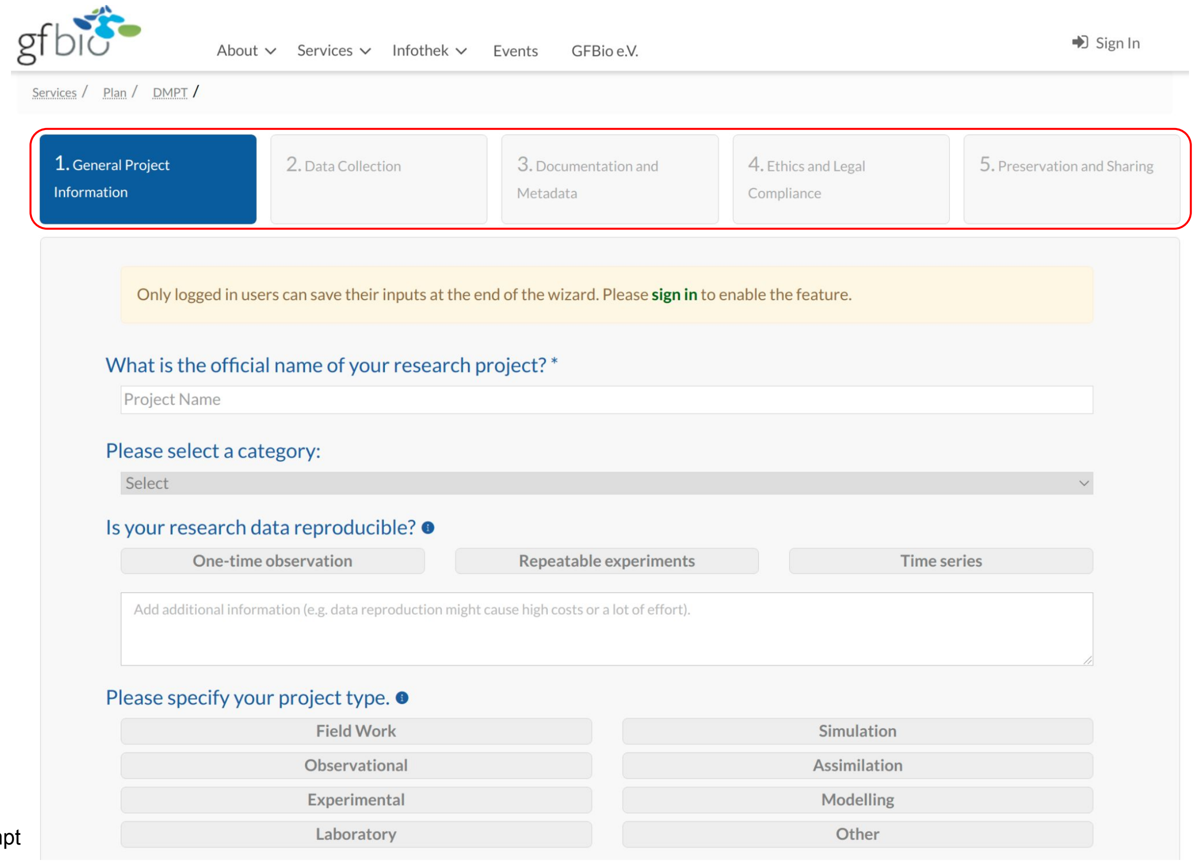
Major migratory routes of birds. Source: birdlife.org

The Data life cycle - Plan

- Identify a study question and formulate a set of hypotheses
- Check existing guidelines, standards and formats (project-specific, institution-wide, domain-wide, etc.)
- Formulate a data management plan (DMP)
- Apply for funding to cover data curation and preservation costs, e.g. at DFG

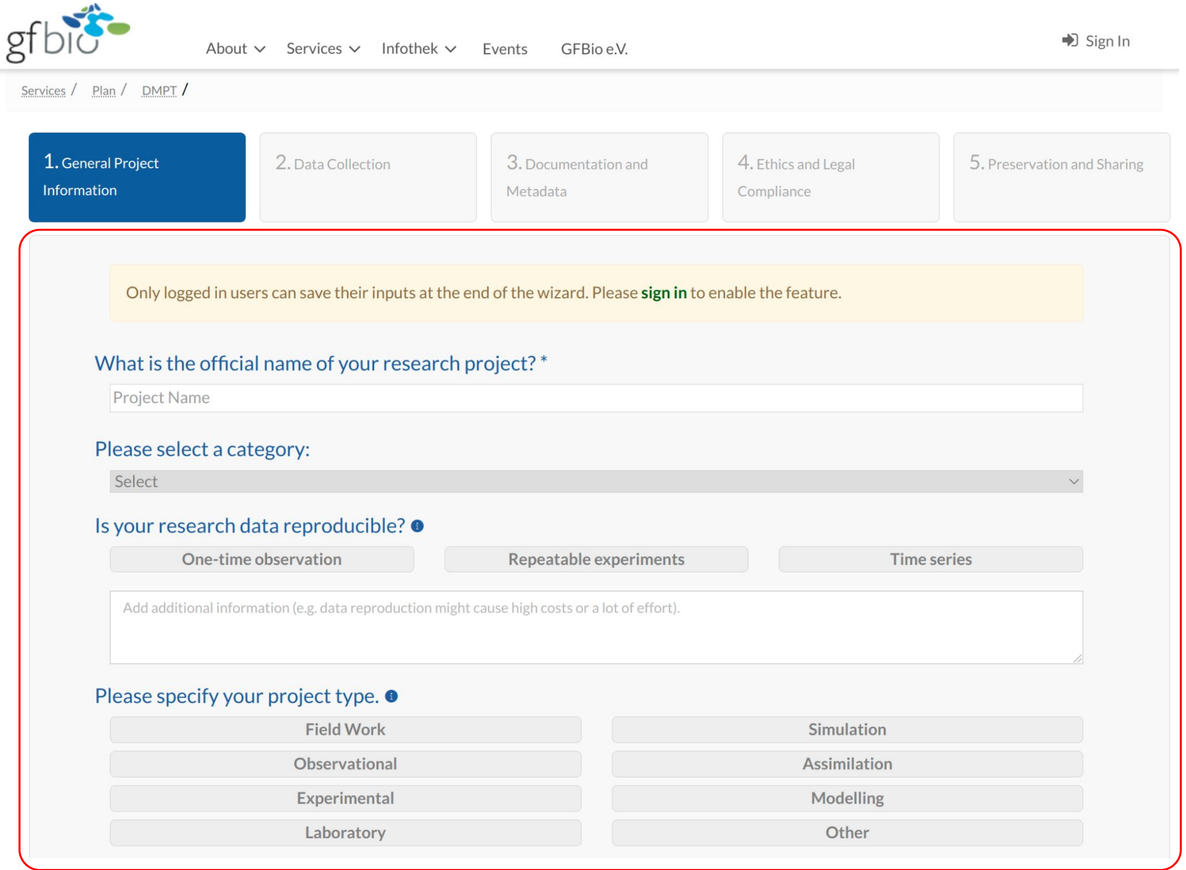


Data management plans



Source: https://www.gfbio.org/dmpt

The data management plan



gfbio

About ⌄ Services ⌄ Infothek ⌄ Events GFBio e.V. Sign In

Services / Plan / DMPT /

1. General Project Information | 2. Data Collection | 3. Documentation and Metadata | 4. Ethics and Legal Compliance | 5. Preservation and Sharing

Only logged in users can save their inputs at the end of the wizard. Please **sign in** to enable the feature.

What is the official name of your research project? *

Project Name

Please select a category:

Select

Is your research data reproducible?

One-time observation | Repeatable experiments | Time series

Add additional information (e.g. data reproduction might cause high costs or a lot of effort).

Please specify your project type.

Field Work | Simulation
Observational | Assimilation
Experimental | Modelling
Laboratory | Other

Source: https://www.gfbio.org/dmpt

Research project - Example DMP (ultra short version)

Data description: This study will use ~600.000 GBIF-mediated occurrence records, monthly climate data from Worldclim 2.0, and functional trait data from Storchová & Hořák (2018). Quality assurance will be carried out in the R statistical programming language (R Core Team 2021).

Documentation and metadata: An ODMAP protocol of the study will be prepared.

Ethical and legal compliance: not applicable

Storage and backup plan: Large data files will be stored locally during the research project. Code versioning will be realized via GitHub.

Preservation: The analysed dataset will be uploaded to a public repository.

Data sharing and publication: All data are publicly available.

Responsibilities: The principle investigator is responsible for all data-related tasks

Resources: No additional hardware resources are needed for this project

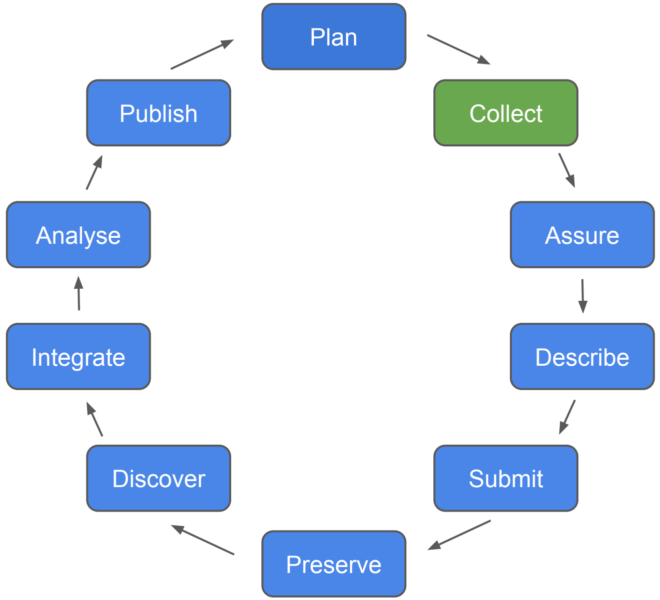
Thematic excursion

Version Control

(switch lectures)

The Data life cycle - Collect

- Set up a collection protocol and collect data in a consistent, systematic manner according to your plan
- Capture and create structured Metadata
- Store your data redundantly (multiple drives in different physical locations)

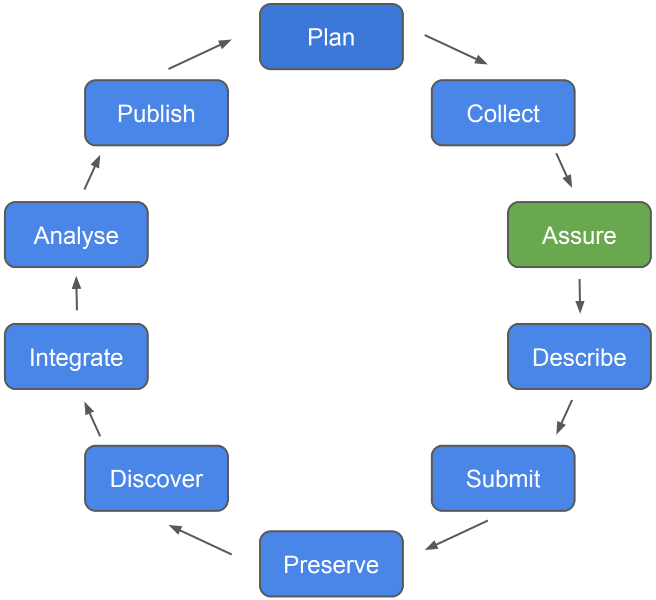


The Data life cycle - Collect

- A few inspirations

The Data life cycle - Assure

- Check for accuracy and consistency of data structure and format
- Detect errors (statistical/graphical analysis)
- Go through the data cleaning process and create a script of it
- Version your data sets
- Document the quality of data by flags, metadata, coding



Source: https://www.gfbio.org/training/materials/data-lifecycle/assure

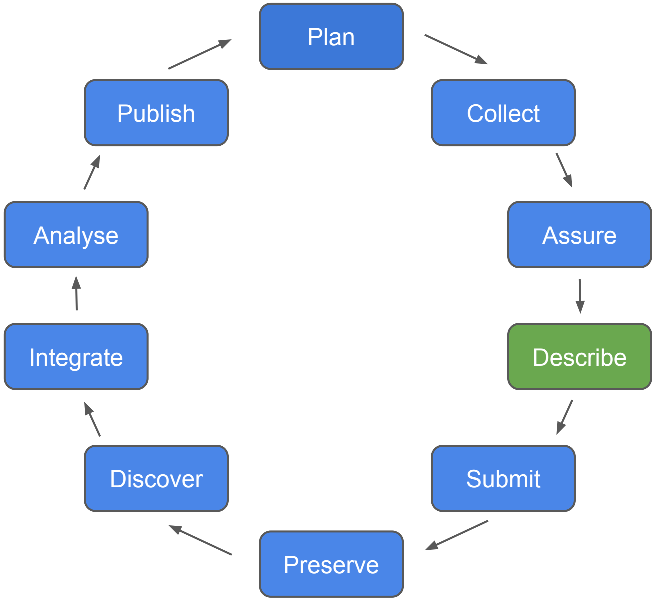
Thematic excursion

Databases

(switch lectures)

The Data life cycle - Describe

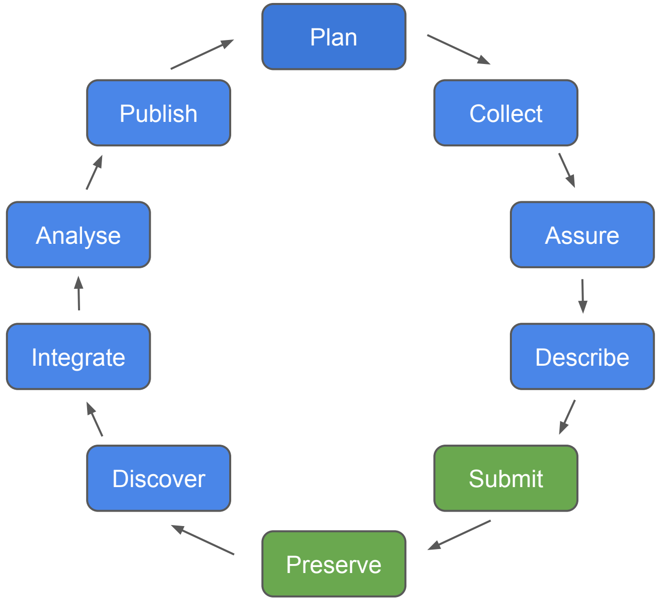
- Metadata should answer the following six questions: **Why** were the data generated? **Who** created the data? **Where** and **when** were the data collected? **What** is the content of the data? How were the data assessed?
- Produce consistent, precise and self-contained metadata
- Ideally, use a metadata standard (e.g. EML, ABCD) or consider converting to a compatible metadata standard at a later point



Source: https://www.gfbio.org/training/materials/data-lifecycle/describe

The Data life cycle - Submit & Preserve

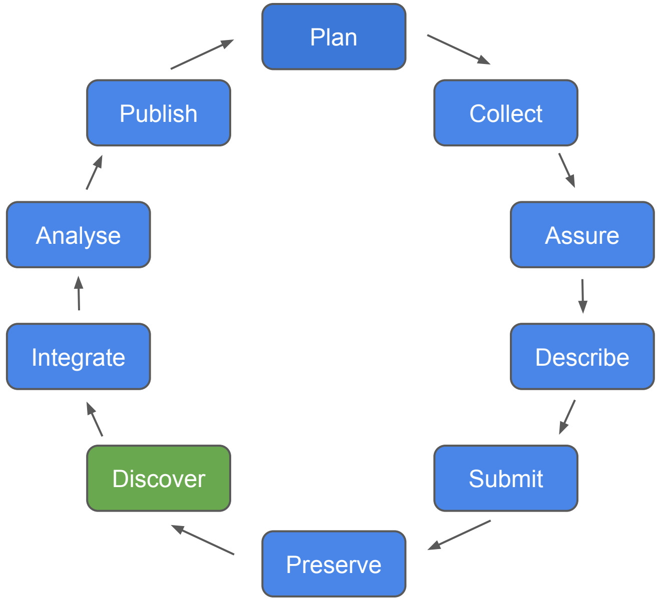
- Upload your data to a long-term archive
- Define availability (publication embargo, access restrictions, etc.)
- Ensure:
  - Accessibility: Data can be retrieved, displayed and used.
  - Authenticity: Data have not been manipulated, substituted or faked.
  - Longevity: Data are re-usable for long-term, independently of software and hardware decay



Source: https://www.gfbio.org/training/materials/data-lifecycle

The Data life cycle - Discover

- Discover suitable datasets to address, expand or verify your research question
- Use appropriate key terms in your search
- Assess suitability through metadata screening, visualization and exploratory analysis
- Check the access requirements and authentication procedures.



Source: https://www.gfbio.org/training/materials/data-lifecycle/discover

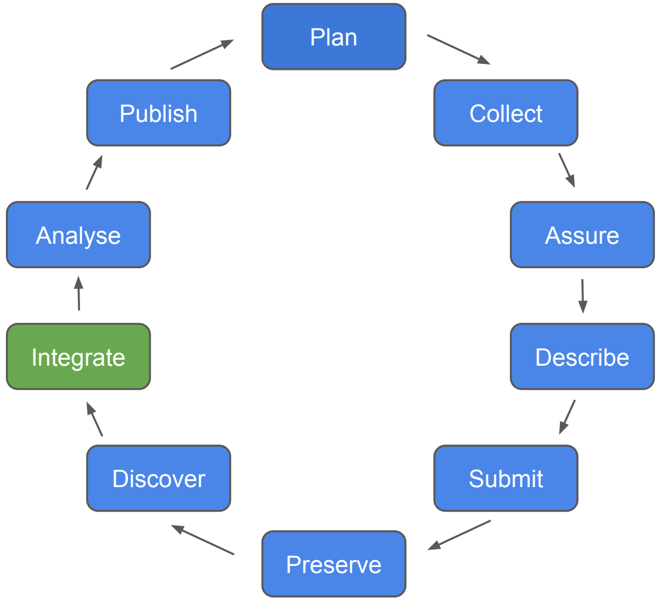
Thematic excursion

Spatial data

(switch lectures)

The Data life cycle - Integrate

- Benefit from best practices in preceding data life cycle stages
- Document integration and analysis workflows
- Check data provenance, avoid duplication
- Verify data and metadata quality
- Cite re-used/integrated data sets accordingly



Source: https://www.gfbio.org/training/materials/data-lifecycle/integrate

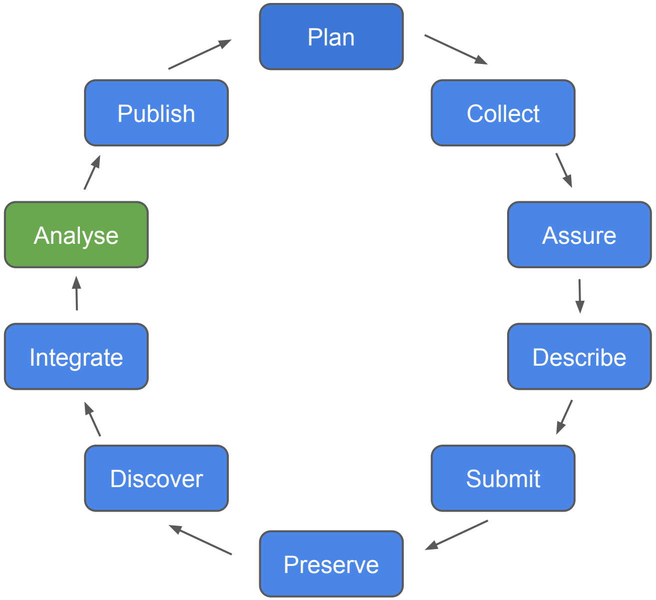
Thematic excursion

Performance

(switch lectures)

The Data life cycle - Analyze

- Think about reproducibility!
- Data analysis with appropriate and accessible tools (data mining, descriptive statistics, graphical maps).
- Careful documentation of analysis workflows

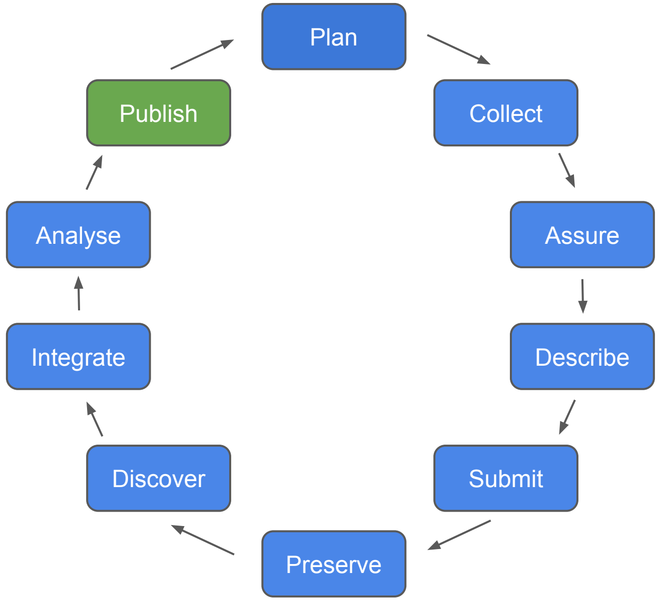


Source: https://www.gfbio.org/training/materials/data-lifecycle/analyze

Analysis and Visualisation: Open discussion

The Data life cycle - Publish

- Prefer journals with open, transparent data policies
- Get a persistent identifier (e.g. DOI) on your submitted dataset.
- Update your archived data sets.
- Impose data embargo if necessary
- Follow community protocols for documentation and metadata publication (e.g. ODD, ODMAP, etc.)



Source: https://www.gfbio.org/training/materials/data-lifecycle/publish

Concluding discussion