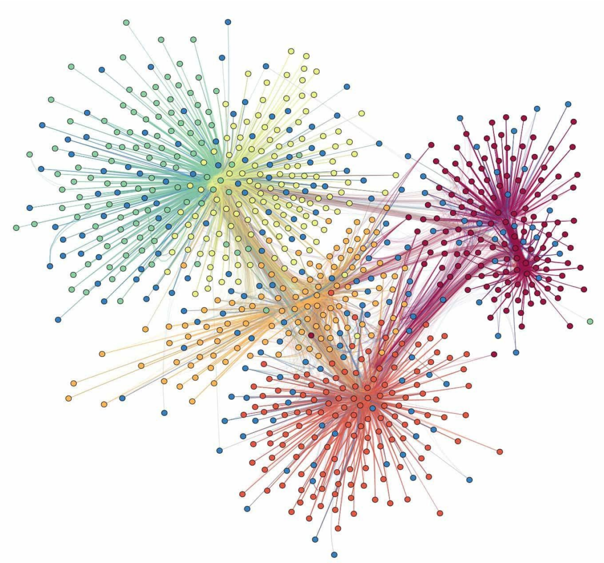
Big Data Ecology

Fundamentals of reproducible and impactful 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

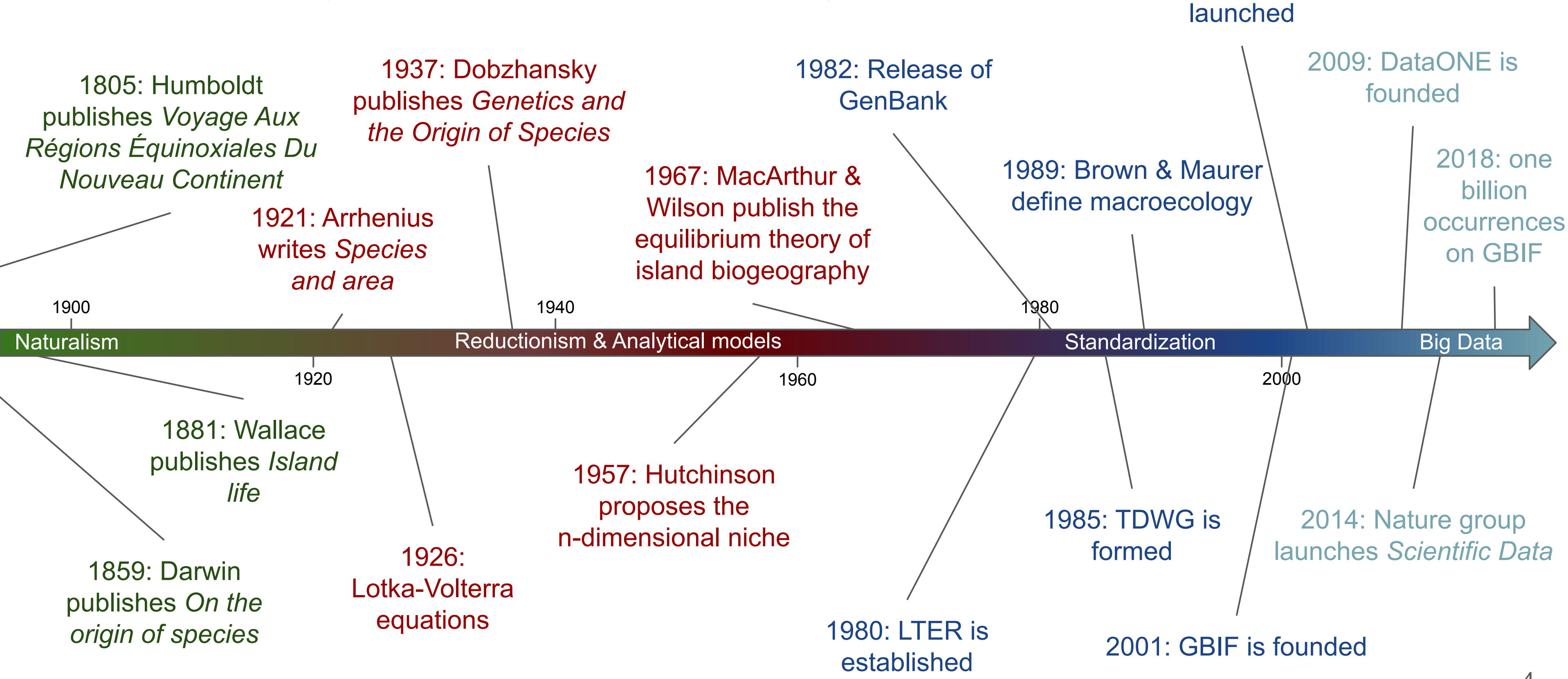
Schedule

Day 1
- Welcome & Introduction
- Standards, Protocols & Concepts
- Version control
- Data bases

Day 2
- Spatial data
- Data processing & computational performance

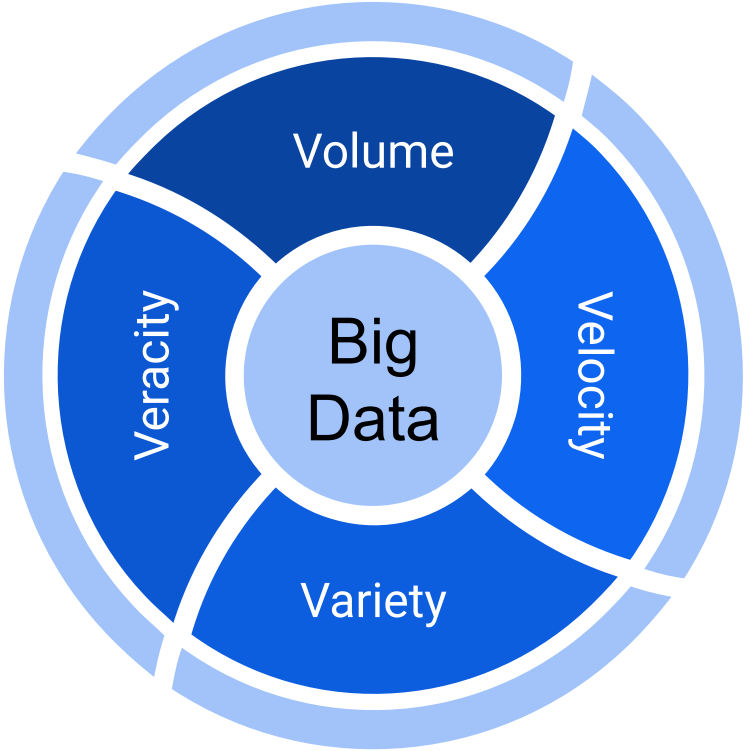
Day 3
- Data analysis & visualization
- Wrap up, concluding discussion & feedback

Biodiversity data - a brief history



What is "Big Data"

Large size of the data



Varying reliability of the data

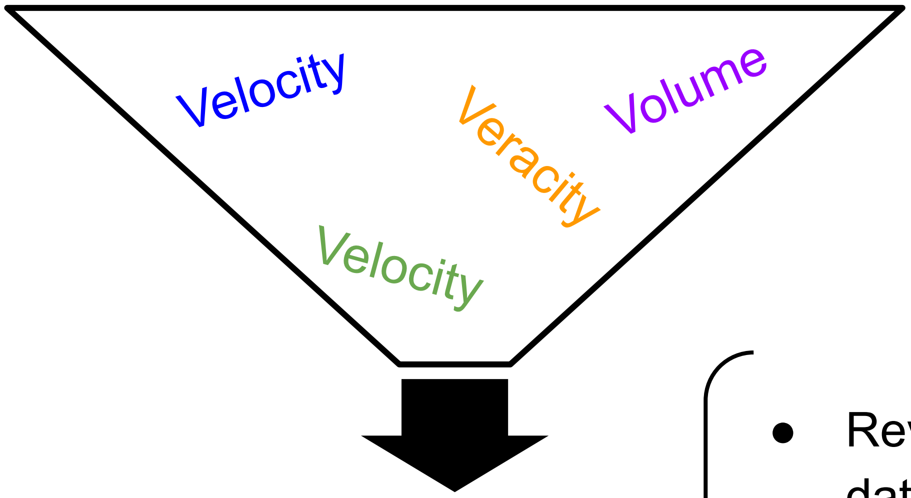
High speed at which data is generated

Many different types of 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”



Value

- 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

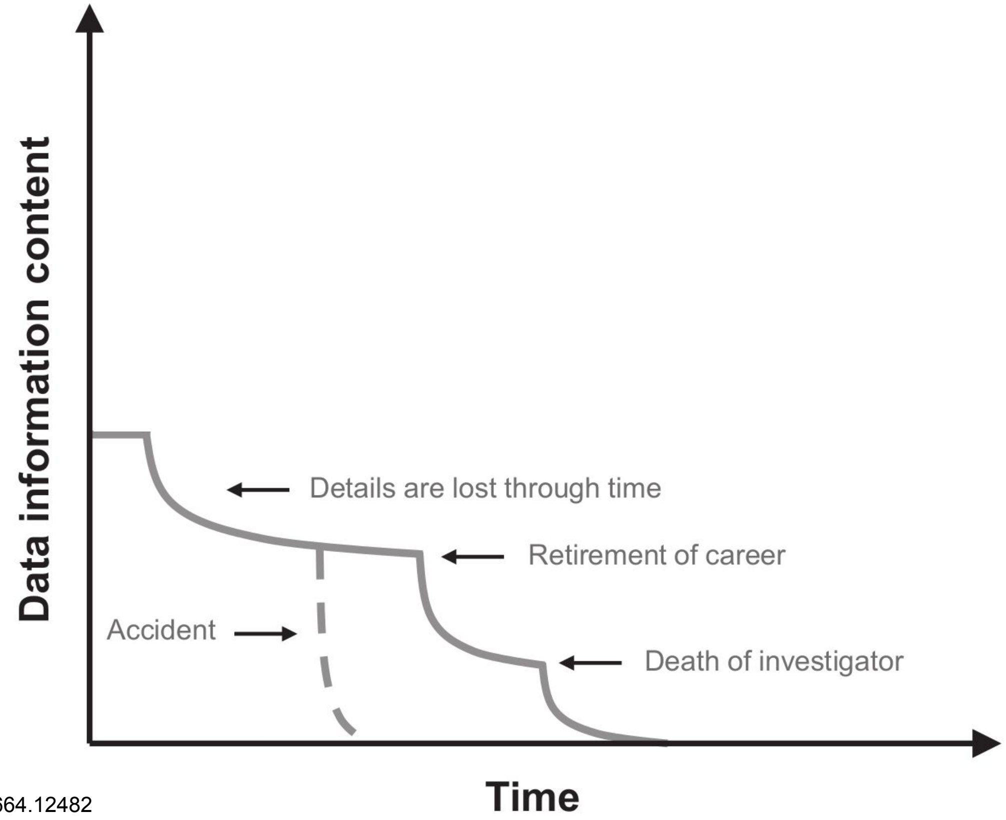
SUBJECT CATEGORIES
» Research data
» Publication characteristics

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/

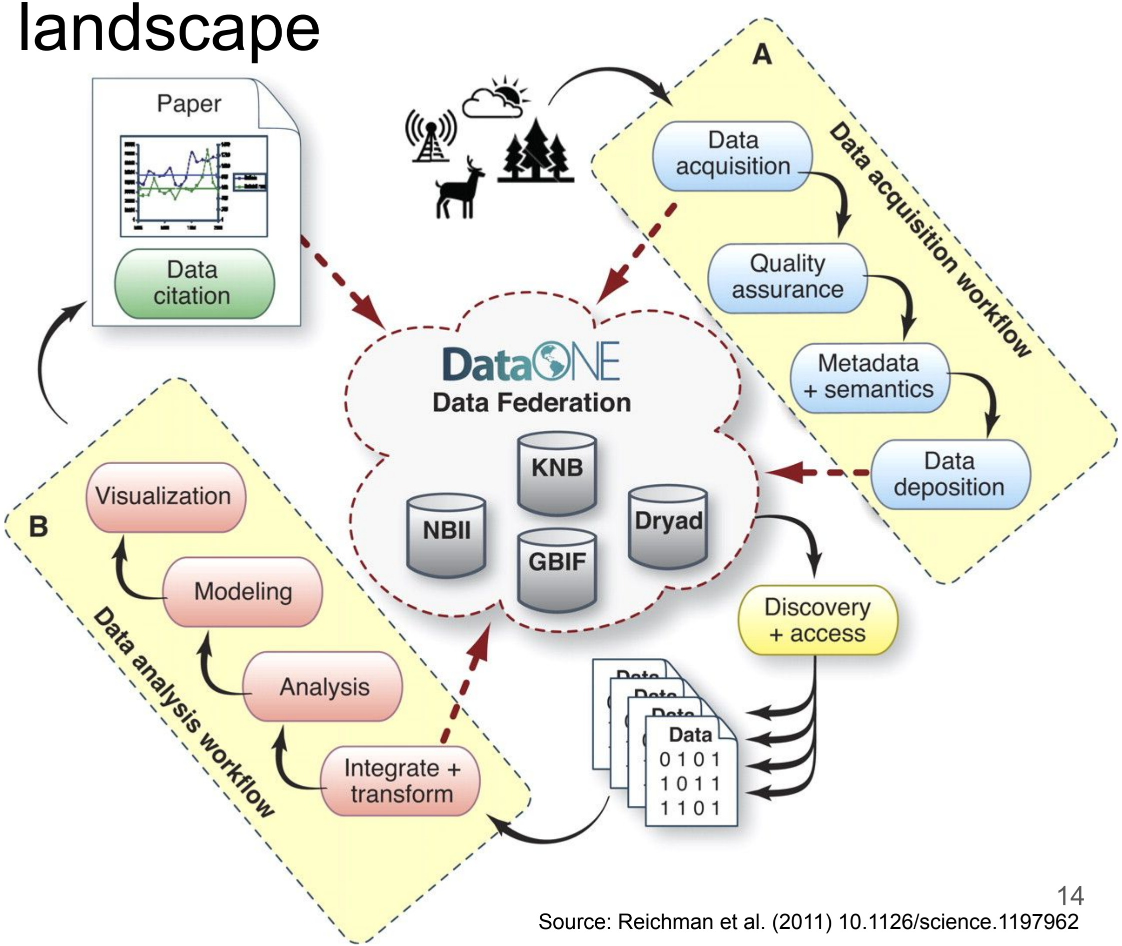
What are key terms mentioned in the FAIR principles

?

A FAIR biodiversity data landscape

Three key components:

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



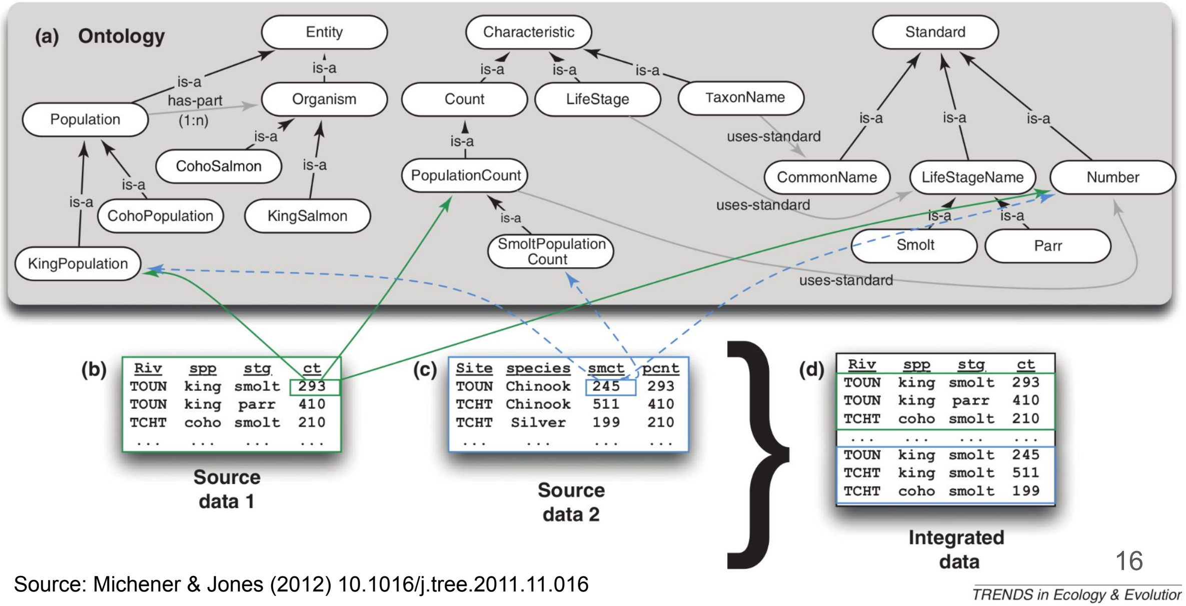
Source: Reichman et al. (2011) 10.1126/science.1197962

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



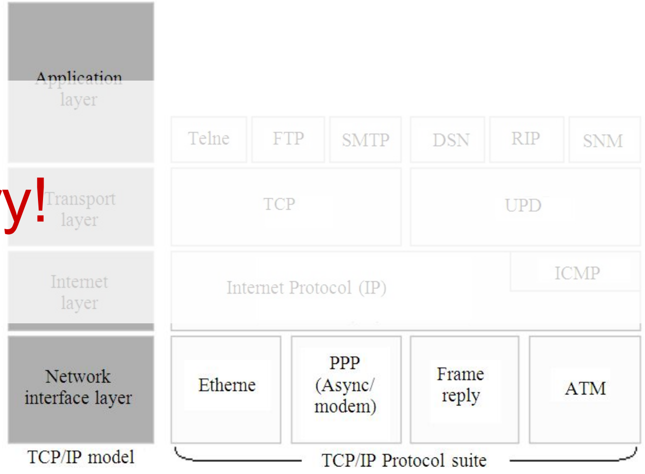
Source: Michener & Jones (2012) 10.1016/j.tree.2011.11.016

A FAIR biodiversity data landscape - Standards & Protocols

International System of Units (SI)

A success story!

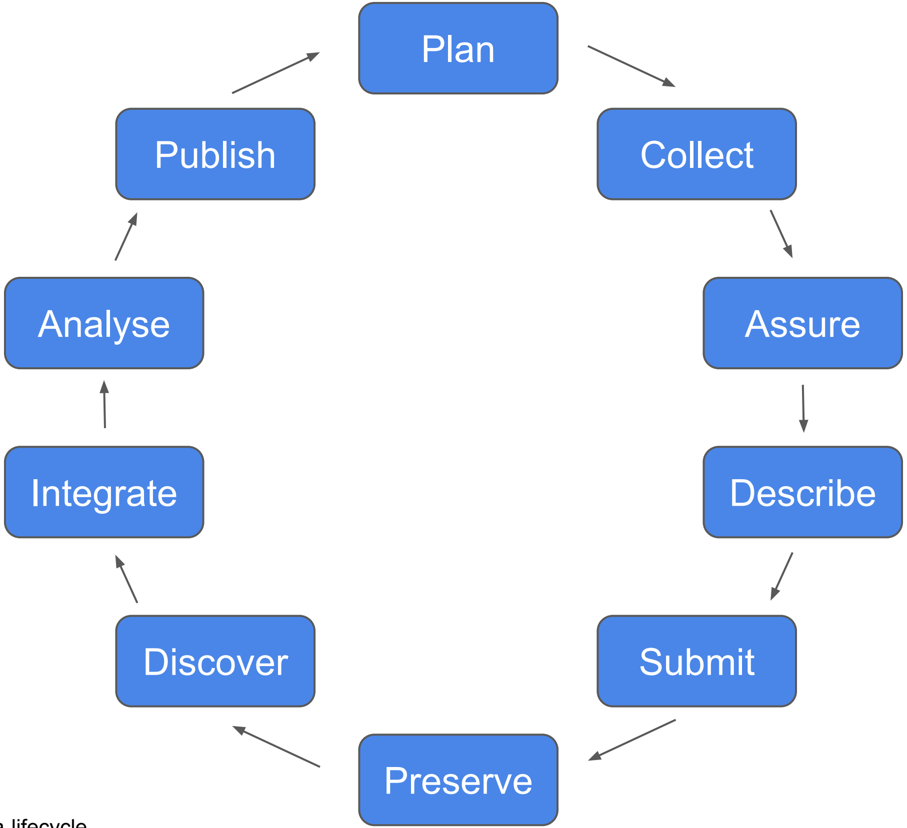
The internet



A FAIR biodiversity data landscape - Federation

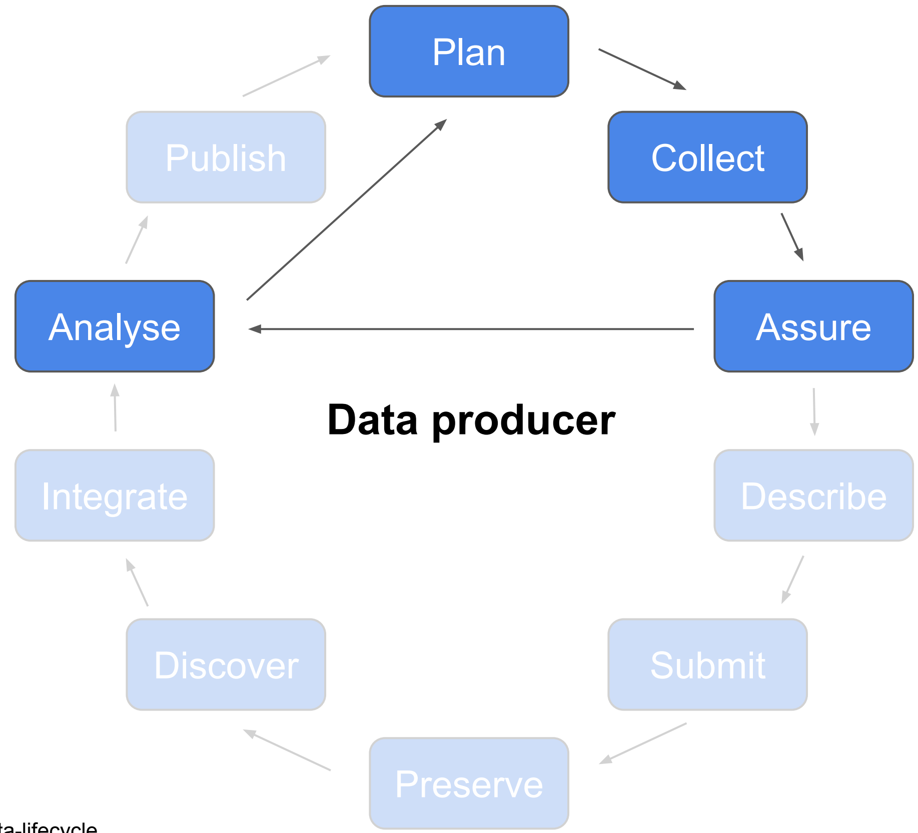
- Distributed management of data by member nodes with proven domain-expertise
- Integration of domain-specific repositories in a *network of databases*
- Search, discovery, access and use of data from different repositories via a common interface
- Advantages:
  - Distributed workload, computation and 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



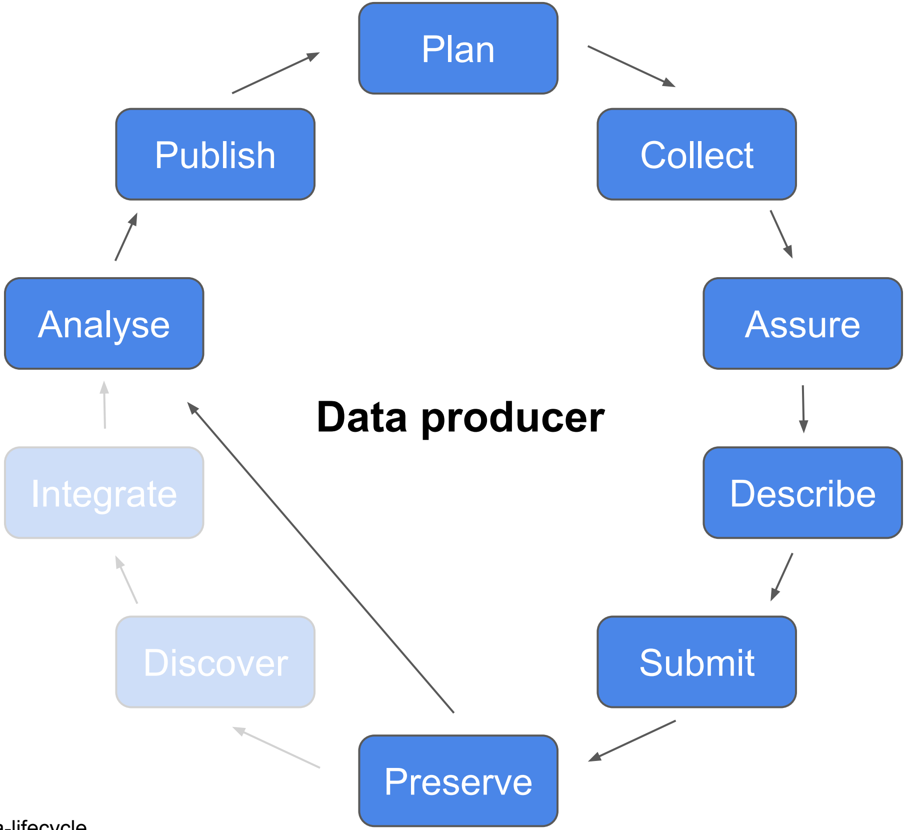
Source: https://www.gfbio.org/training/materials/data-lifecycle

A FAIR biodiversity data landscape - The data life cycle



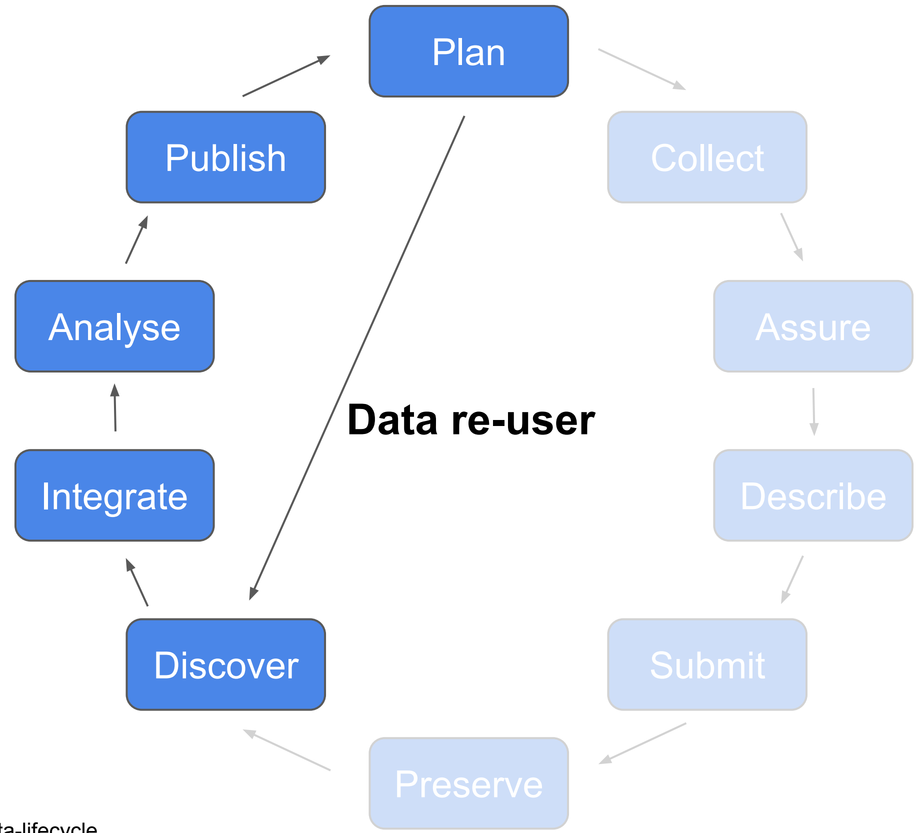
Source: https://www.gfbio.org/training/materials/data-lifecycle

A FAIR biodiversity data landscape - The data life cycle



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

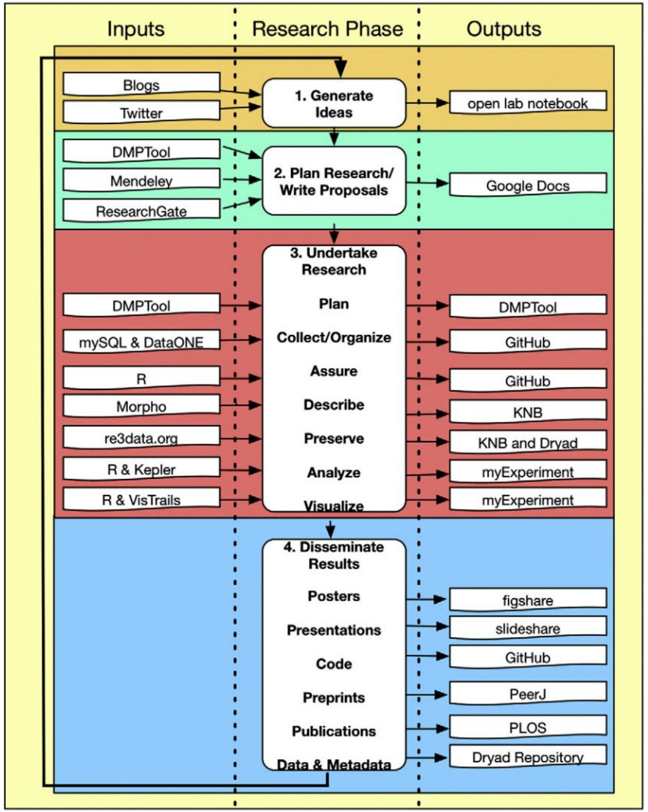
A FAIR biodiversity data landscape - The data life cycle



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

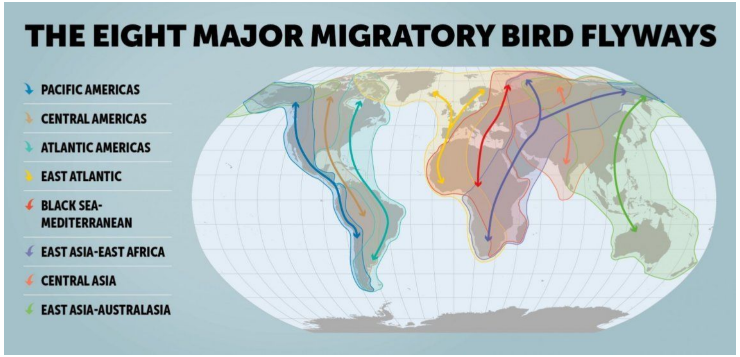
Aim: (1) Explore the imprint of seasonal bird migration in digitally available occurrence records

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

Study taxon: Harriers (*Circus* Lacepede 1799)

Study region: Sweden

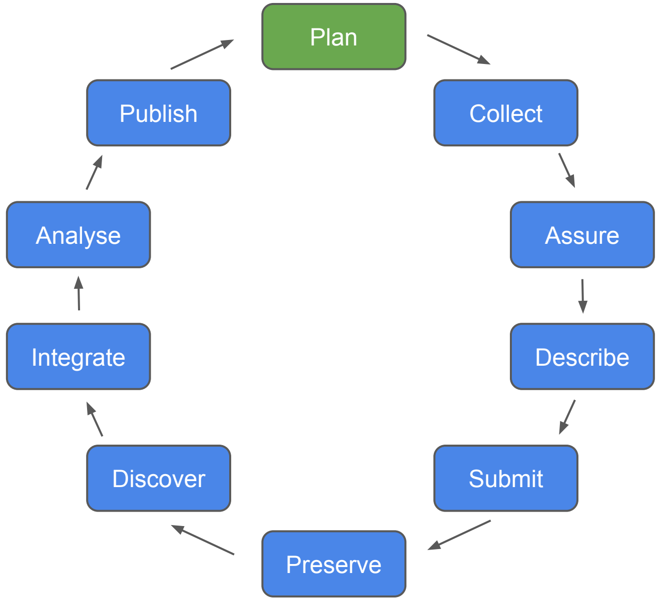
Circus cyaneus in flight. Source: wikimedia commons



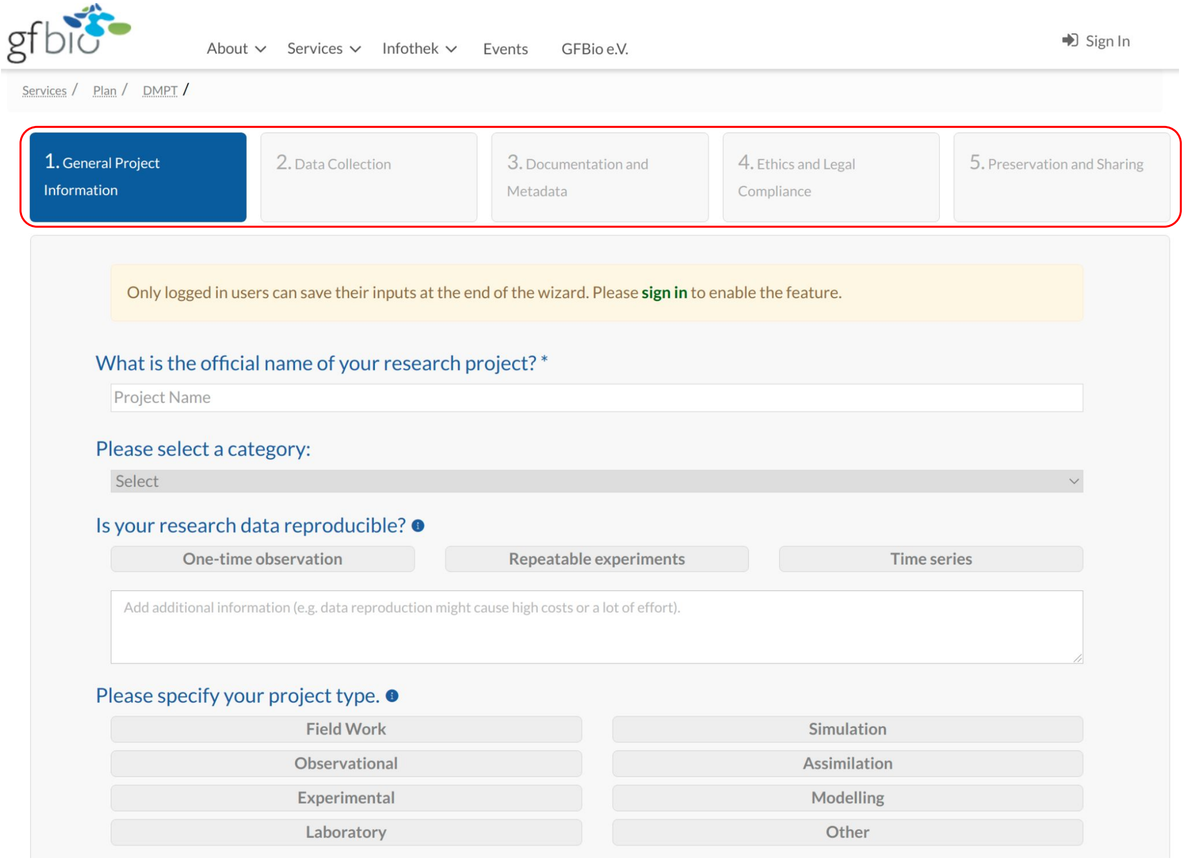
Major migratory routes of birds. Source: birdlife.org

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

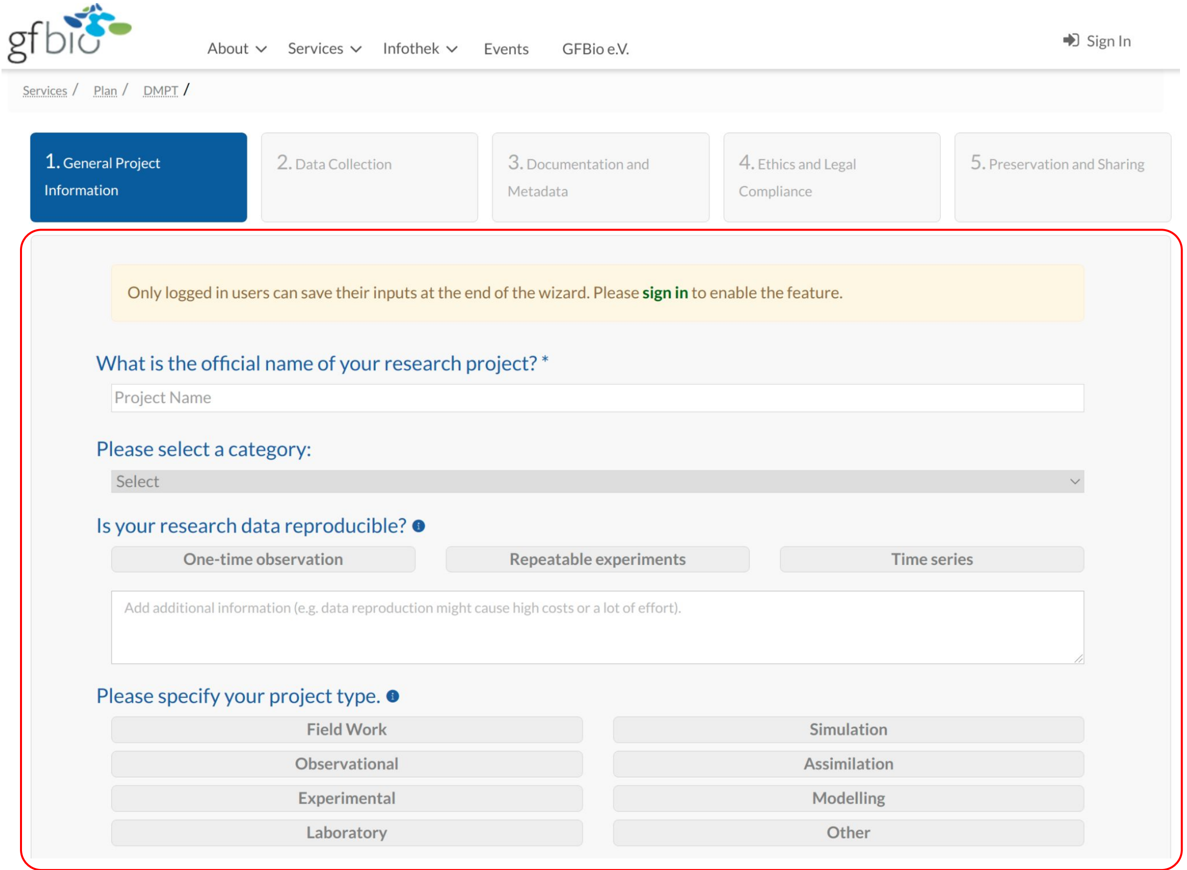


Data management plans



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

The data management plan



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 (you!) is responsible for all data-related tasks

Resources: No additional hardware resources are needed for this project

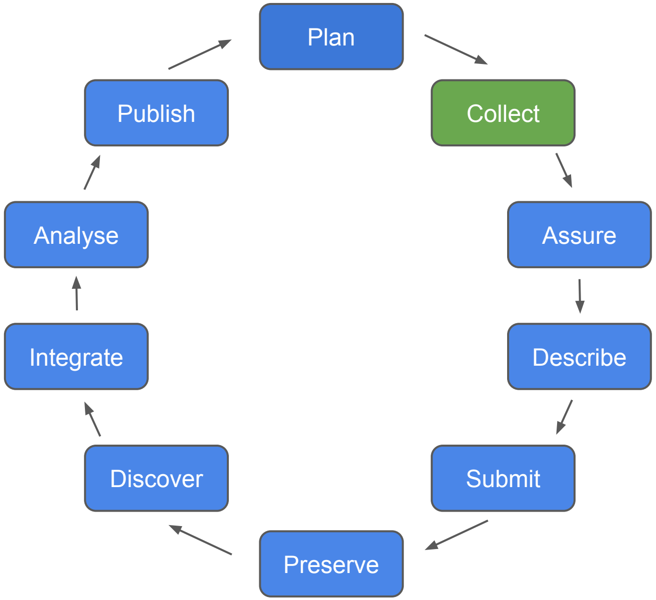
Thematic excursion

Version Control

(switch lectures)

The Data life cycle - Collect

- 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)



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