Data Management for Transparent Research

Serena Bonaretti

https://sbonaretti.github.io/

BiGCaT Science Café

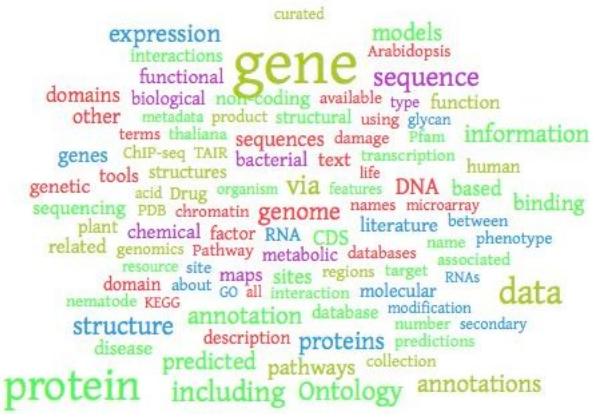
2nd May 2019

CC BY NC SA

Data in life sciences

- Scope, size, and complexity of datasets has increased dramatically over the last 10-20 years
- Traditionally, the knowledge of how to manage this data is often limited to "information managers"
- In response to new funding requirements, researchers are asked to be part of data management

(Griffin et al. 2019)

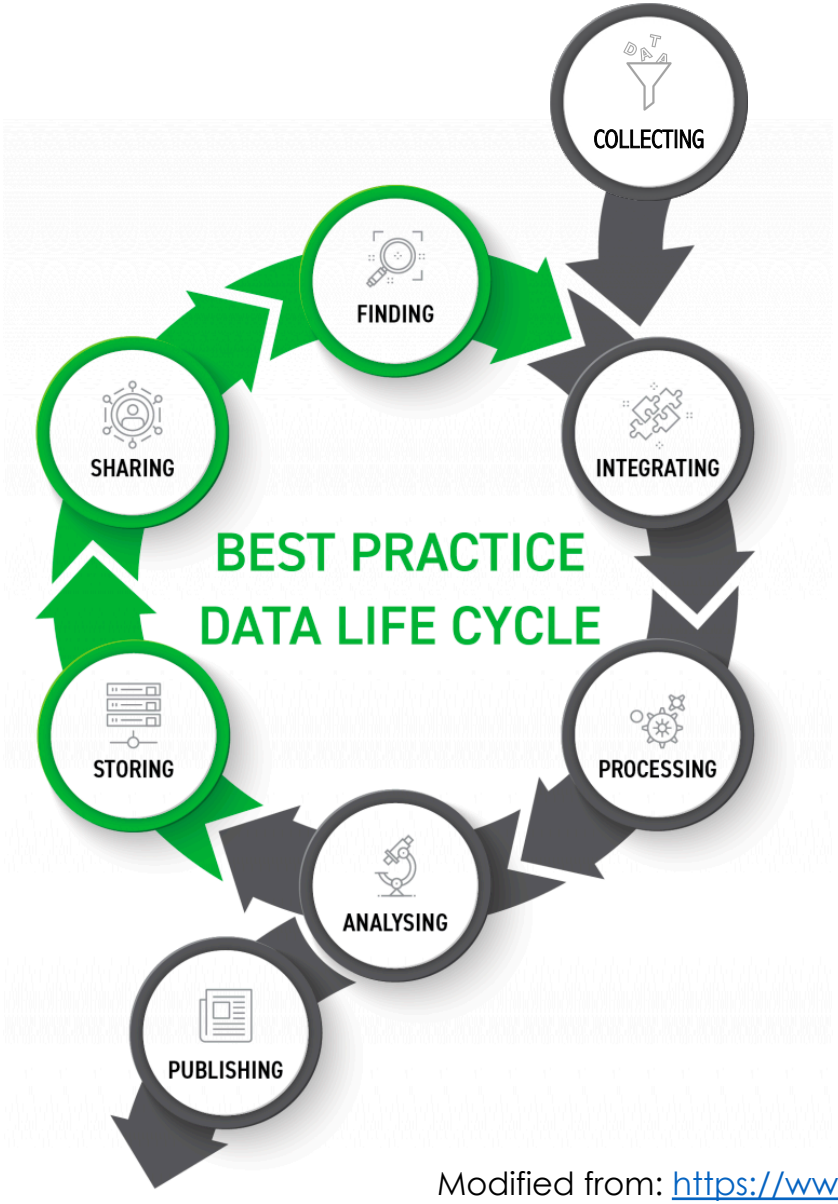


Why do we need to manage data?

- We want to be able to **reuse** data produced by ourselves and by other researchers

- "We are losing data at a rapid rate, with up to 80% unavailable after 20 years" (Griffin et al. 2019)

Data life cycle in research

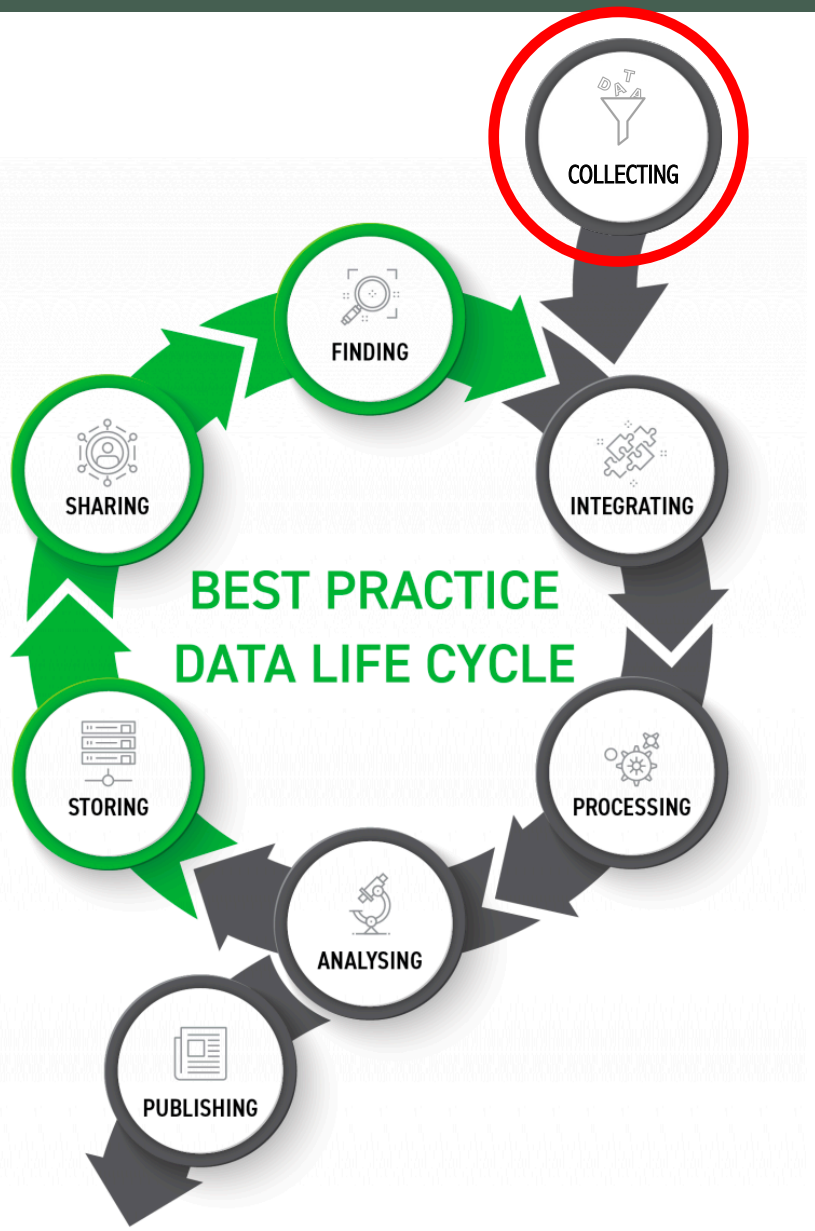


Steps for "traditional" linear view of research data

Steps for data reusability

Modified from: https://www.embl-abr.org.au/data/ and Griffin et al. 2019

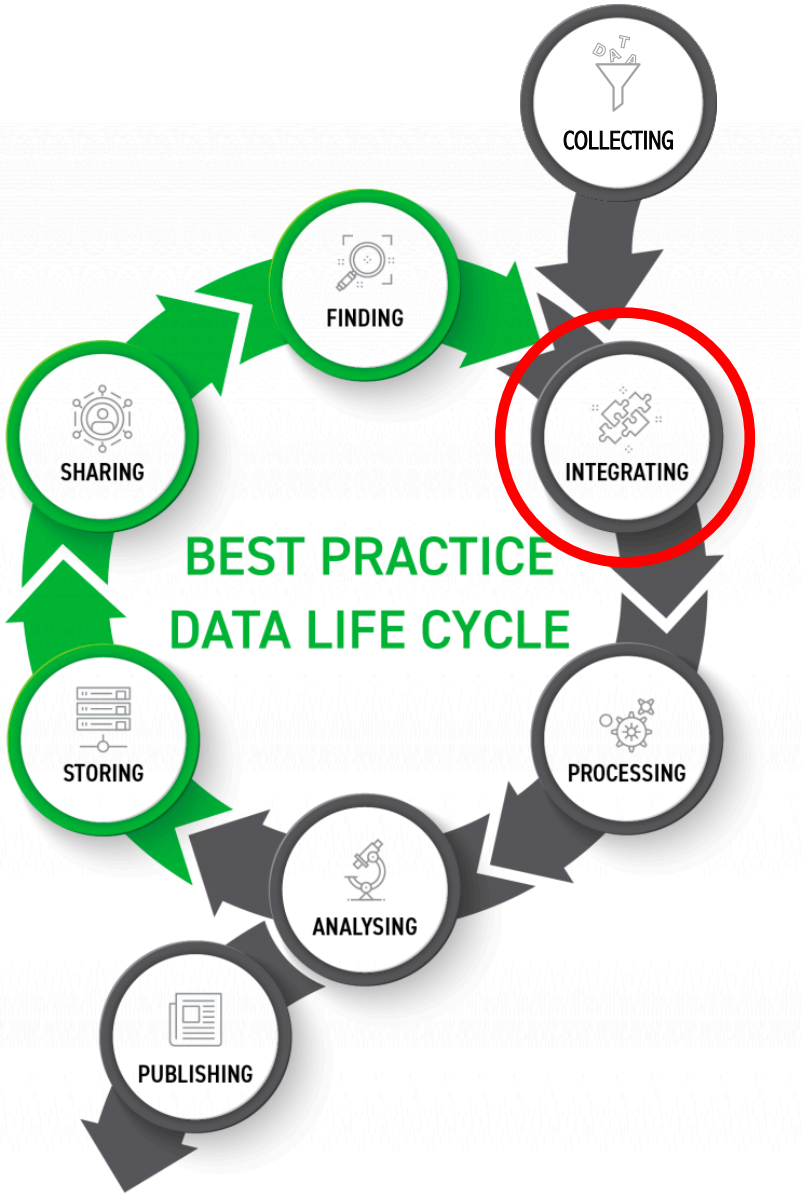
Collecting



- Repositories are indexed in registries of repositories (e.g. re3data, FAIRsharing, etc.)
- Data need to have accurate *metadata*
  - E.g. Use of vocabulary from ontologies (= set of categories that define objects and the relationships among them)

Griffin et al. 2019

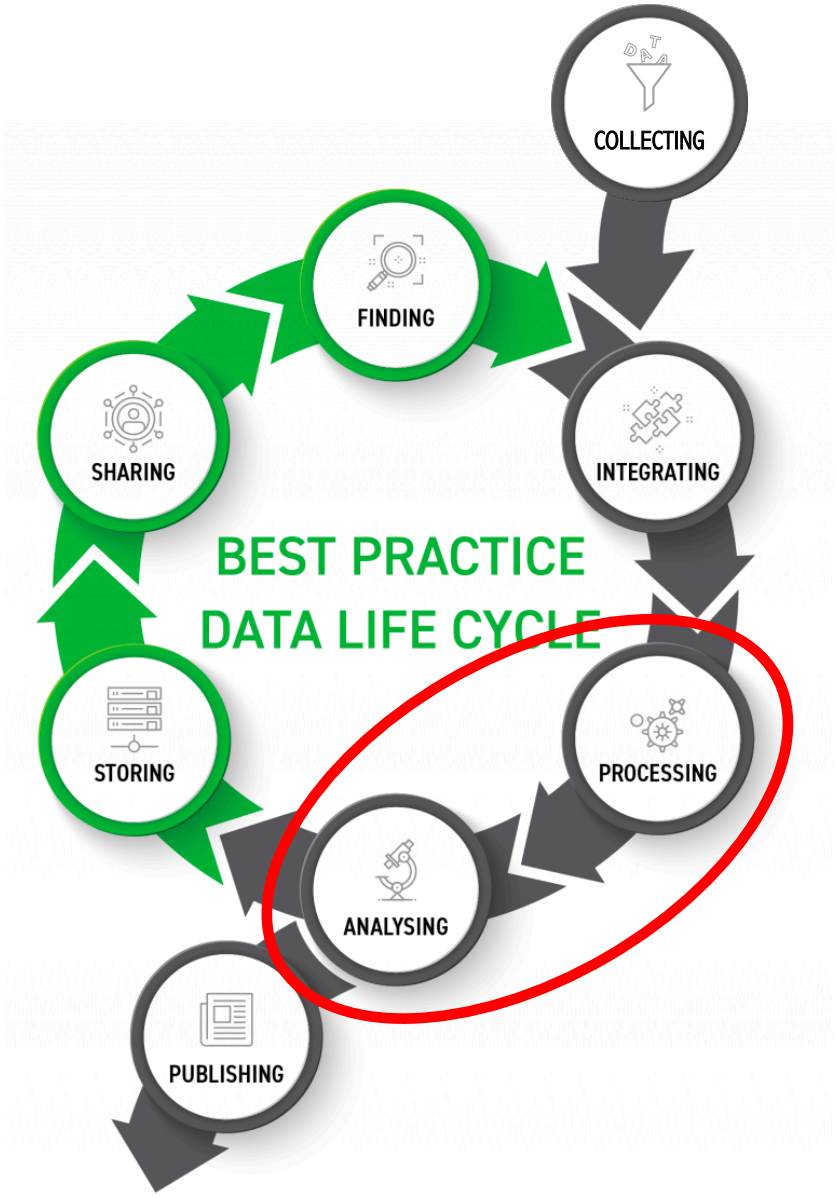
Integrating



- Main issue: Standardization
  - Data can have different formats
- Possible solutions:
  - Linked Data and Semantic Web
  - Format converters

Griffin et al. 2019

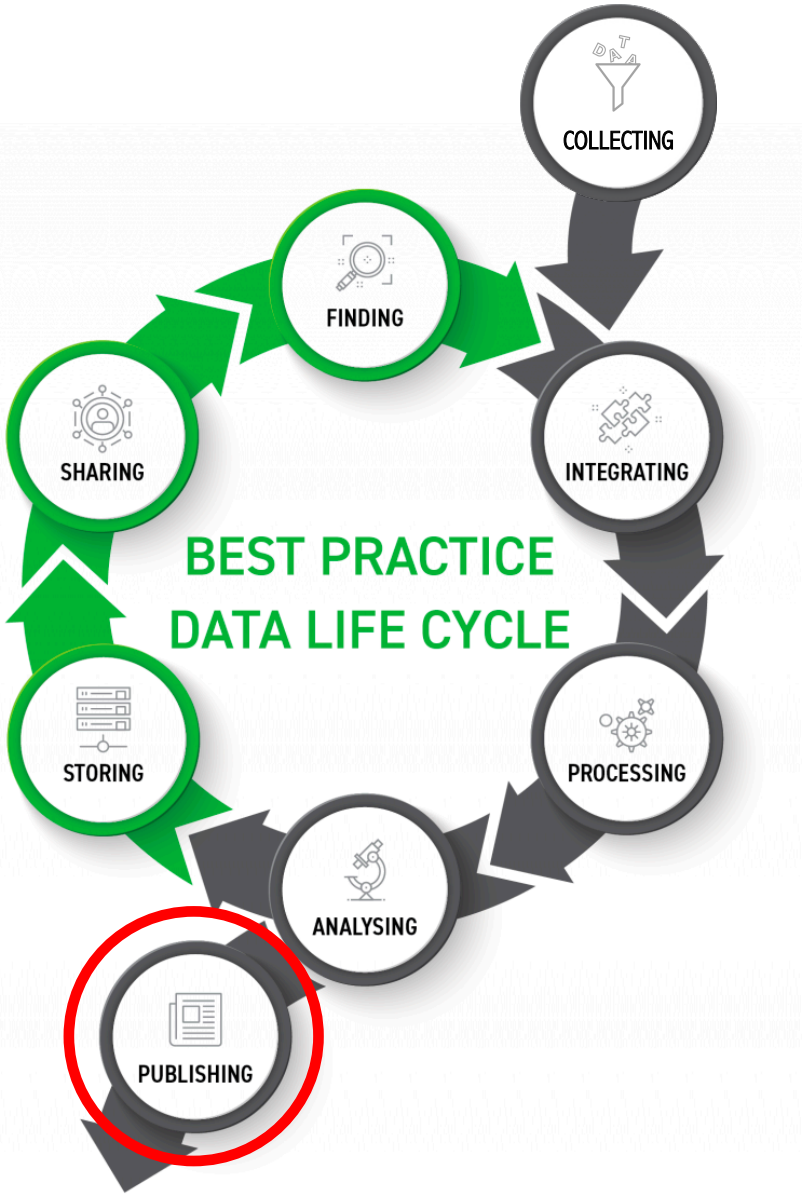
Processing and Analyzing



- Processing and Analyzing are crucial for computational reproducibility and assessment of quality
  - Use of electronic notebooks with dependences
  - Virtual machines containing the whole computational environment (e.g. Docker)
- In the lab: Electronic laboratory notebooks
  - E.g. LabTrove, BlogMyData, Benchling, etc.

Griffin et al. 2019

Publishing

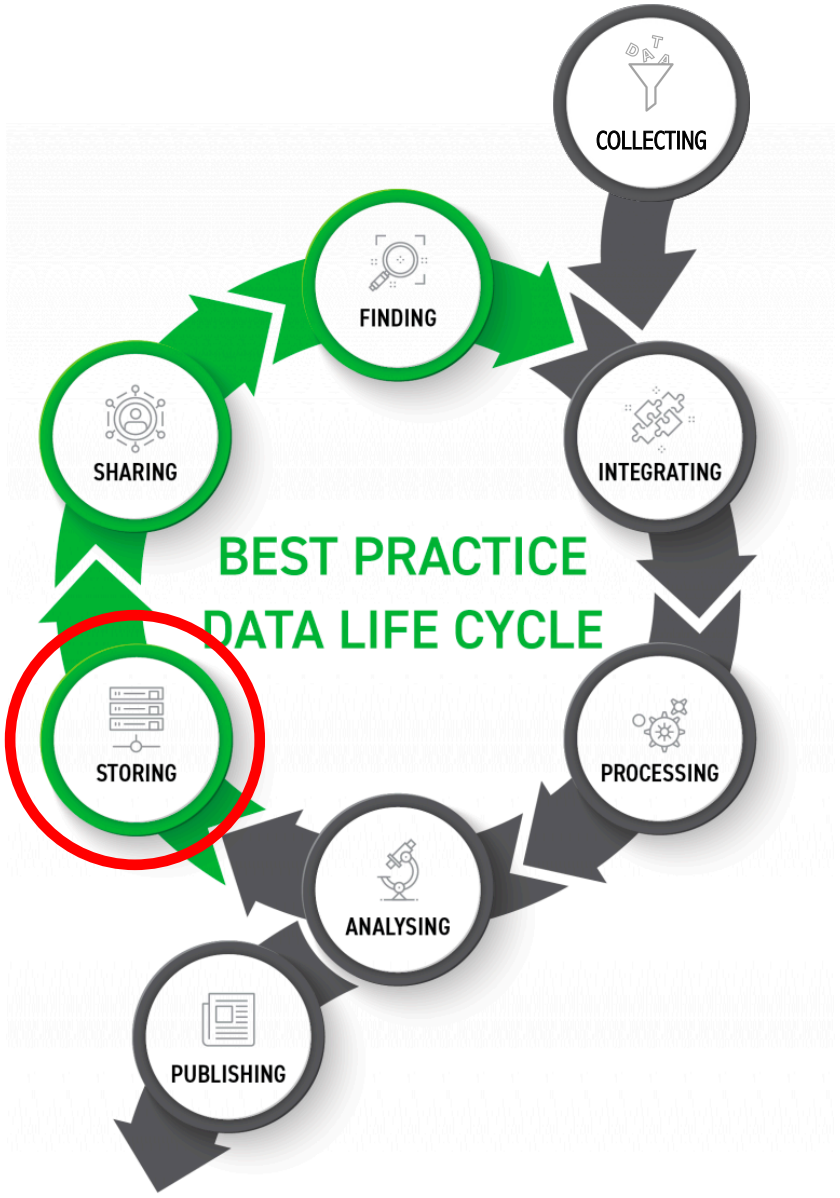


- Journals require more and more to share both raw data and derived data
  - Issues: Ethical constraints and commercially-relevant data

- New journals that publish papers describing datasets
  - Authors can get credit for labor-intensive and expensive data collections
  - E.g. Data descriptor in scientific data, Data note in gigascience, etc.

Griffin et al. 2019

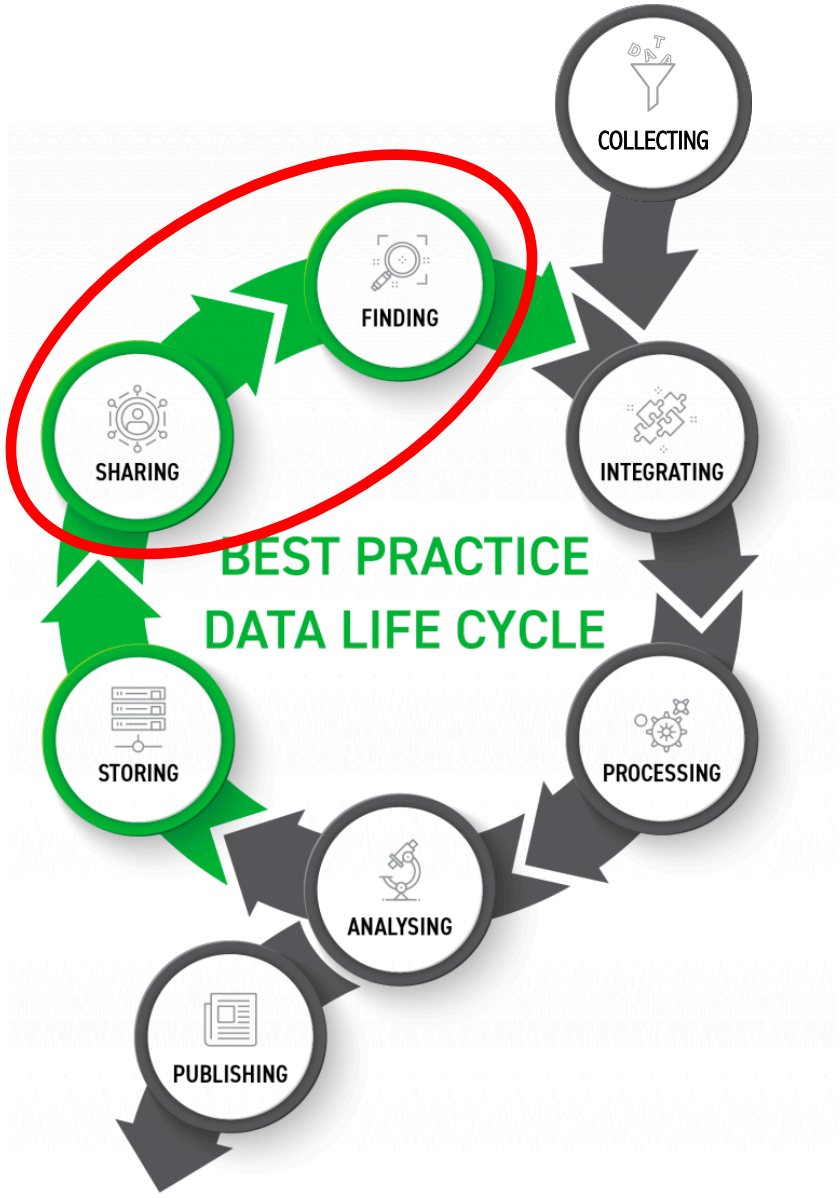
Storing



- Local data storage during collection, integration, processing and analysis
  - E.g. Dropbox, Google Drive, local servers
- Data repositories for sharing
  - E.g. Zenodo, Figshare, domain-specific repositories

Griffin et al. 2019

Sharing and Finding

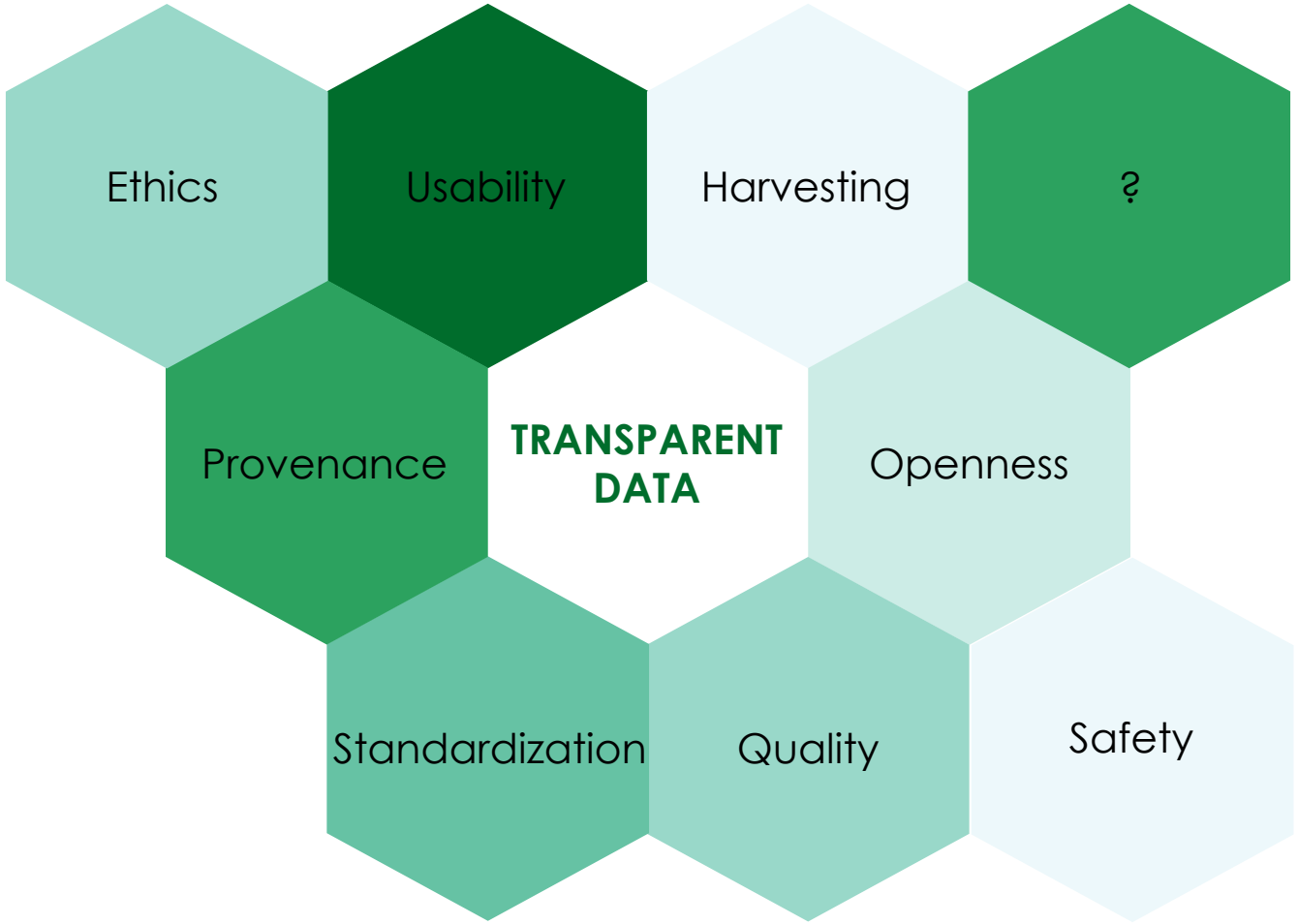


- Characteristics of *open* datasets:
  - Data are hosted in a public repository
  - Data are accompanied by descriptive metadata
  - Data have unique identifiers (e.g DOI)

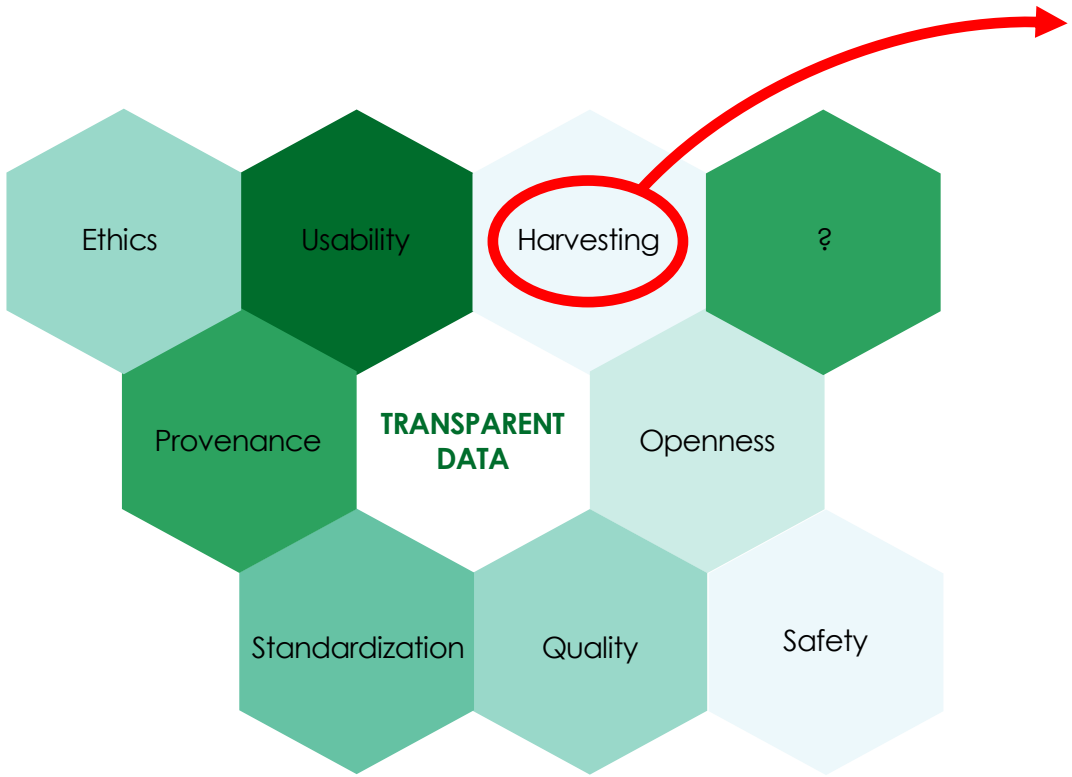
- Issues:
  - Data quality
  - Ethical constraints

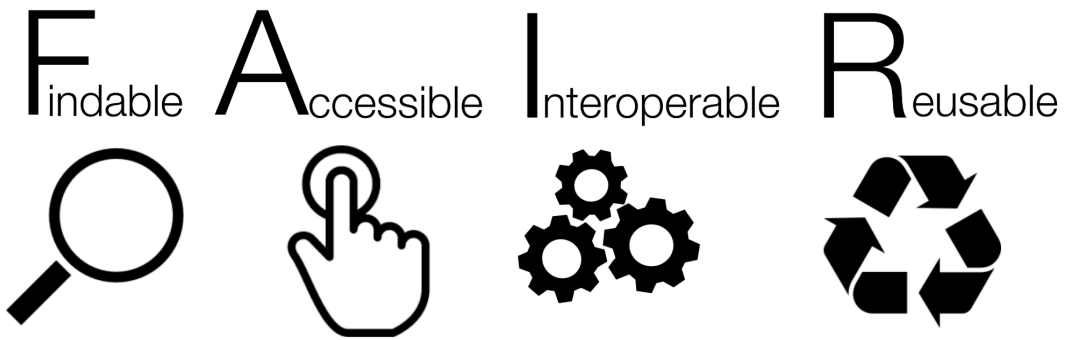
Griffin et al. 2019

Topics in data management



The FAIR guiding principles





To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be 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

To be 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

To be Reusable:
R1. meta(data) are richly described with 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 detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Wilkinson et al. 2016

Academic discussion about FAIR

- What FAIR **is**
  - It "refers to a set of principles, focused on ensuring that research objects are reusable" (Mons et al. 2017)
  - "They deliberately do not specify technical requirements" but "They describe characteristics and aspirations for systems and services" (Mons et al. 2017)
  - *My interpretation:* They are principles to organize metadata in such a way that *machines* can harvest data for us

- What FAIR is **not**
  - It is not a standard, it is not specific to life sciences, it is not equal to semantic web (Mons et al. 2017)
  - It does not imply open data (Mons et al. 2017)
  - *My interpretation*: It is limited to data harvesting, i.e. it does not provide guidelines for format standardization, data quality, etc.