Small, thick, and slow

Towards an Open and FAIR data culture in the Humanities

Daniel Paul O’Donnell
University of Lethbridge

IIT Gandhinagar
December 18, 2019

DOI (this version): 10.5281/zenodo.3581520
DOI (latest version): 10.5281/zenodo.3581519

About this paper

- Going to be speaking of how data are used in the humanities and implications for infrastructure design
    - How infrastructure currently interacts with typical humanities research practices
    - Why humanities researchers have been slow to adopt such infrastructure
    - How this infrastructure can be adapted to support (and improve) humanities research *without requiring it to abandon its primary features/strengths*
        - “Small” — focussed on very small number of data points or sets
        - “Thick” — involves intense curation and analysis of these few data
        - “Slow” — the same data points can be subject to years (generations) of subsequent, alternate, and supplementary analysis

About this paper

- Important to recognise that I’m dealing in generalities
  - Not all humanities data are small or “representational” in focus
  - Not all humanities work is about thick description
  - Not all humanities work is about reworking old material
- But much is and these are the ones that are least well catered to in current infrastructure

About me

- Traditionally trained medieval philologist and textual critic
- Means history of “big” and small data techniques
  - Thesis (1996) was analysis of (unpublished) database of textual variation in the Old English poetic canon
    - Letter-by-letter differences in about 20 poems surviving in more than one copy from the pre-conquest period
  - Later (2005) did 100,000 word edition of 9-line *Cædmon’s Hymn* (s. viii)
  - Now working on 5 object “edition” of the cross in pre-conquest England
- But
  - Coming from a textual/linguistic/literary approach
  - Focus on “editing” (i.e. the development and publication of “Primary Source” material — mediated representational data)

Traditionally, humanists resist speaking of data

- “Primary sources” = Texts, artifacts, objects of study
  - Can be originals (i.e. the artifact itself)
  - More often mediated and contextualised in some way (i.e. an edition, transcription, or similar)
- “Secondary sources” = Works of other scholars (often based on “Primary sources”)
- “Readings” (1) = Passages, extracts, quotations for interpretation or support
- “Readings” (2) = Interpretation, the end product of research (literary study)

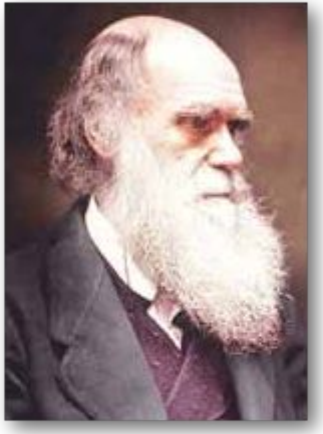
Traditionally, humanists resist speaking of data

- These definitions are highly contingent
  - “Primary source” in one context can be the “secondary source” in another (and vice versa)
  - Or simultaneously “Primary” and “Secondary” (e.g. a critical edition)
- Also hard to constrain
  “[a]lmost any document, physical artifact, or record or human activity can be used to study culture” and arguments proposing previously unrecognised sources (“high school yearbooks, cookbooks, or wear patterns in the floors of public places”) are valued acts of scholarship”

(Borgman 2007)

How does data work in other fields?

- Resistance makes sense, because Humanities data are different from other forms of data
- In other domains, “data” (“given things”) are often more properly “capta” (“taken”): generated through experiment, observation, and measurement, then analysed
- Think about Darwin and his work in the Galapagos Islands
  - What are his data?



How does data work in other fields?

- Resistance makes sense, because Humanities data are different from other forms of data
- In other domains, “data” (“given things”) are often more properly “capta” (“taken”): generated through experiment, observation, and measurement, then analysed
- Think about Darwin and his work in the Galapagos Islands
  - What are his data?

The finches?

How does data work in other fields?

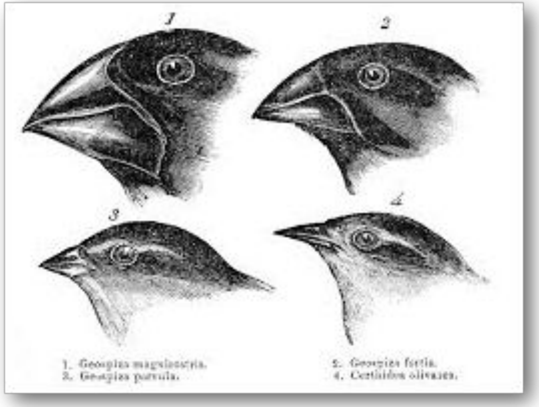
- Resistance makes sense, because Humanities data are different from other forms of data
- In other domains, “data” (“given things”) are often more properly “capta” (“taken”): generated through experiment, observation, and measurement, then analysed
- Think about Darwin and his work in the Galapagos Islands
  - What are his data?

The notes about the finches?

How does data work in other fields?

“represent[ation of] information **in a formalized manner** suitable for communication, interpretation, or processing” (NASA 2012);

“the facts, numbers, letters, and symbols **that describe** an object, idea, condition, situation, or other factors” (NRC 1999)



The finches.

But in the humanities?

- Can be both “data” and “capta”, but very often “data”:
  - Very specific and often provisional (i.e. **small**);
  - Dialogic in nature — defined and changed by the interpretative context (i.e. **thick**);
  - Frequently revisited to reanalyse or recontextualise (i.e. **slow**).
- Not produced so much as found.

But in the humanities?

- Can be both “data” and “capta”, but very often “data”:
  - Very specific and often provisional (i.e. **small**);
  - Dialogic in nature — defined and changed by the interpretative context (i.e. **thick**);
  - Frequently revisited to reanalyse or recontextualise (i.e. **slow**).
- Not produced so much as found.

Usually the Finch.

But in the humanities?

- Can be both “data” and “capta”, but very often “data”:
    - Very specific and often provisional (i.e. **small**);
    - Dialogic in nature — defined and changed by the interpretative context (i.e. **thick**);
    - Frequently revisited to reanalyse or recontextualise (i.e. **slow**).
- Not produced so much as found.

Usually the Finch. Sometimes the notes.

But in the humanities?

- Can be both “data” and “capta”, but very often “data”:
  - Very specific and often provisional (i.e. **small**);
  - Dialogic in nature — defined and changed by the interpretative context (i.e. **thick**);
  - Frequently revisited to reanalyse or recontextualise (i.e. **slow**).
- Not produced so much as found.

Usually the Finch. Sometimes the notes. And sometimes Darwin.

But in the Humanities?

- Some evidence:
  1. Humanities “data,” unlike science “capta,” are almost always practically and theoretically **non-rivalrous**.
     - Humanities researchers rarely have an incentive (or capability) to prevent others from accessing their raw material.
     - 200 years of Jane Austen studies based on five main pieces of data.
  2. No traditional methods around the preservation of those “facts, numbers, letters, and symbols **that describe** an object, idea, condition, situation, or other factors.”
     - Nobody cares if your notes are well-preserved, unlike lab books.

The “Digital Humanities” don’t change this

- DH adds to this basic distinction, but doesn’t change it:
  - We can now have “capta” (intermediate “observations” extracted algorithmically to form large data sets that then require interpretation)
  - We can now work across complete historical or geographic corpora:
    - E.g. all known nineteenth-century English periodicals; every surviving tract from the U.S. Civil War
  - We can do deductive, hypothesis-driven work
  - Introduces issue of reproducibility:
    - Since deductive work based on algorithmically derived data-sets can be checked (for completeness, opposing examples, etc), it can be more important to publish method.
    - People might care about your notes.

The “Digital Humanities” don’t change this

- But the distinction between “capta” and “data” is not teleological — *DH (and especially not “big data” DH) is not the perfection of the Humanities*
    - “Big data” (“big capta”) DH is not better than “small data” Humanities — it allows different kinds of questions and approaches to questions
    - Not all DH is “big capta” — you can answer traditional humanities questions with the assistance of computation
- Sometimes (many times) “big capta” simply misses the point
    - Intensive curation and analysis of small data sets remains a major function of humanities research (small)
    - Dialogic definition of data remains a major method (thick)
    - Revisiting data to reflect current concerns remains a core purpose (slow)

Why does this matter?

- Although much humanities research is (appropriately) “small, thick, and slow,” it is also, in theory, useful for “big capta” work
    - Collectively, traditional humanists produce a lot of *very* high quality data
        - *Intensely* curated datasets and data points;
        - Broadly compatible with each other (i.e. each generation re-edits and reconsiders the canon; reconsiders historical events, etc.);
- If we could find a way to capture the value of this traditional data in a way that would allow them to be reused,
    - We’d have extremely useful material to repurpose
    - We’d be maximising the benefit of the traditional work that has been done on it

FAIR data

- In many sciences, there is a technically similar opportunity (though it involves a very different purpose):
    - Traditionally, scientists have not been good at publishing their data — they’ve published the analysis and conclusions (i.e. the relevant bits)
    - Reasons have included
        - Lack of fora: no means to distribute data;
        - Lack of will: publishing data means exposing real world messiness
        - Lack of reward: data are not “first class research objects” (credit is meaningless)
    - This hinders both reproducibility (can’t easily check your method) and development of new science (can’t build on your results)

FAIR data

- FAIR (Findable, Accessible, Interoperable, Reusable) data and data citation principles are attempt to address this by establishing data as “first class research objects” — i.e. objects for which scientists can get credit
- Goal is to standardise and formalise the way in which data is **published** in order to ensure its entry into the scientific record and reuse.
- Encourage data publication, solving the forum problem
- FAIR data can be
  - Found and accessed by others without negotiation
  - Used in new use-cases or to reexamine old ones

FAIR data

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.

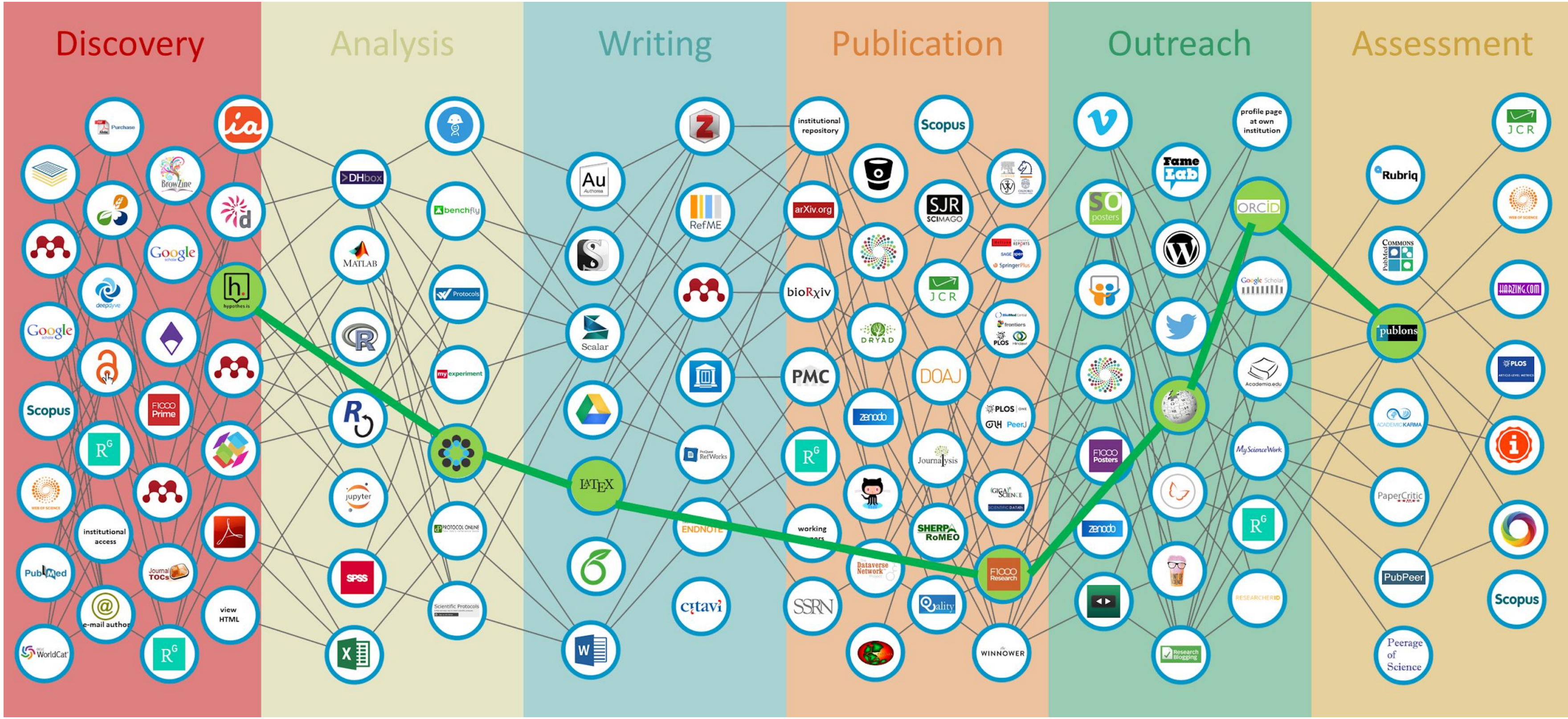
A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A2 metadata are accessible, even when the data are no longer available.

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.

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.

FORCE11 (2014): http://bit.ly/IITGN-FAIR

FAIR Data



Bosman and Kramer 2016

FAIR Humanities?

- You might think this would work well for Humanists
    - Publication of Data as “First class research object” is inherent in several traditional strands of humanities research
        - Editing
        - GLAM
        - Concordances and Corpora

FAIR Humanities?

- But FAIR small data is by-and-large uneconomical (and uninteresting) for small data researchers — whether digital or traditional
    - Reproducibility is still generally not a priority
        - Disagreement about the role of women in 19th C factories more likely a difference of emphasis or interpretation than falsification
        - You don’t need my precise dataset to critique my argument
    - Even data publishers focus on small, contextualised, datasets
        - e.g. an edition of Jane Austen’s *Pride and Prejudice* is intended to support secondary work on that novel — not work on novels generally

FAIR Humanities?

- The features that are required for reuse require (in essence) a separate, standalone, publication
  - Deposit in repository
  - Standardised metadata
  - (Potentially) loss of key interpretative context and information

- None of this is (necessarily) required for success of original publication
- Much of it (may) require additional work and costs that detract from without improving the original research
  - Unlike STEM, reproducibility is simply generally not an issue of importance

FAIR Humanities

- Instead of a modular workflow as in Bosman and Kramer 2016, in the Humanities, (data) publication (and processing) still tends to be done on a project-by-project basis
    - Individual websites
    - Larger “clubs” that you have to join (Europeana)
- The purpose in each case is what you might call “local” publication:
    - Making the data accessible for project-specific purposes, without concern for Interoperability or Reusability
- Has longer term implications as well:
    - Projects die (URLs rot)
    - Not machine readable

The case of manuscript photography

- Since the mid-1990s, there have been hundreds if not thousands of digital editions published of medieval and renaissance texts.
- Almost all of these contain high quality digital photographs of the original artifacts, often with very detailed, research-based expert commentary and analysis (transcriptions, bibliographic and other descriptions, etc.)
- Represents, *in theory*, a potentially huge, extremely rich, dataset for new cross-project work
    - Automatic scribe identification
    - Dating training sets
    - History of the Book

The case of manuscript photography

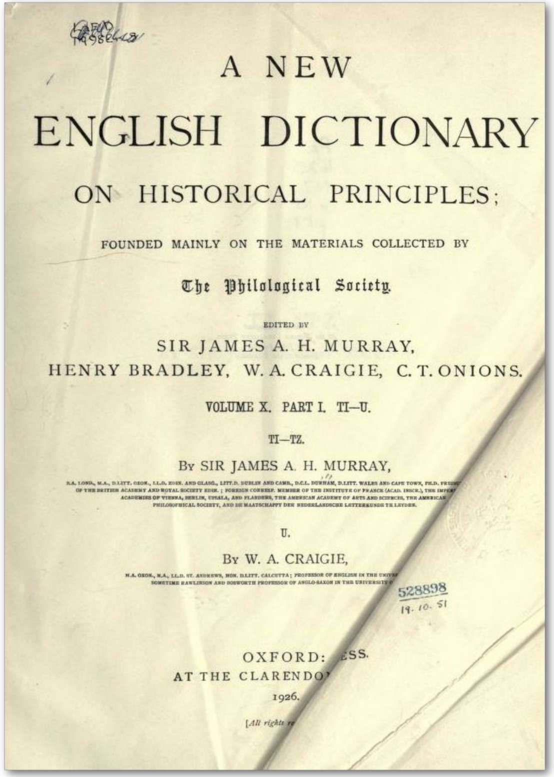
- Because the purpose of these photographs has been to support the contextual analysis and/or supply users with representations of the individual objects in question, very few are easily recovered or used by machines:
    - Few/no standards for metadata, APIs, etc
    - Very few explicitly connected to expert description
    - Relationship to other images and publication status not machine readable
- Result is a lost opportunity to create a “big capta” dataset of thickly described data from hundred of individual “small data” projects
- Reason was that it was in nobody’s interest to contribute to the Commons

So what to do?

- The solution to this is to accept the traditional nature and use-case involved in the production and consumption of Humanities research data
    - I.e. recognise that FAIR must accommodate the small, thick, and slow as easily as it does the big stand-alone examples from STEM
- That means that we have to either
    - Work within the traditional Humanities research workflow
    - Encourage traditional Humanities researchers to work within ours
- **As long as FAIR data publication means, in essence, publishing small, thick, and slow data twice (once in context and once without), we will never fully reap the benefit of these important and potentially huge cultural datasets**

We’ve been here before

- The New English Dictionary provides a non-digital model for this
  - Based on “historical principles” (i.e. definitions from and supported by historical quotations)
  - Massive crowd-sourced big-data collection, involving thousands collecting 1.8 million quotation slips from thousands of books prepared by generations of authors, scholars, and publishers (i.e. small data datasets)
  - In essence, an analogue version of what we want to do digitally



A NEW

ENGLISH DICTIONARY

ON HISTORICAL PRINCIPLES;

FOUNDED MAINLY ON THE MATERIALS COLLECTED BY

The Philological Society.

EDITED BY

SIR JAMES A. H. MURRAY,
HENRY BRADLEY, W. A. CRAIGIE, C. T. ONIONS.

VOLUME X. PART I. TI—U.

TI—TZ.

By SIR JAMES A. H. MURRAY,

B.A. LOND., M.A., D.LITT. OXON., LL.D. EDIN. AND GLASG., LITT.D. DUBLIN AND CAMB., D.C.L. DURHAM, D.LITT. WALES AND CAPE TOWN, PH.D. FREIBU
OF THE BRITISH ACADEMY AND ROYAL SOCIETY EDIN.; FOREIGN CORRESP. MEMBER OF THE INSTITUTE OF FRANCE (ACAD. INSCR.), THE IMPER
ACADEMIES OF VIENNA, BERLIN, UPSALA, AND FLANDERS, THE AMERICAN ACADEMY OF ARTS AND SCIENCES, THE AMERICAN
PHILOSOPHICAL SOCIETY, AND DE MAATSCHAPPY DER NEDERLANDSCHE LETTERKUNDE TE LEYDEN.

U.

By W. A. CRAIGIE,

M.A. OXON., M.A., LL.D. ST. ANDREWS, HON. D.LITT. CALCUTTA; PROFESSOR OF ENGLISH IN THE UNIVE
SOMETIME RAWLINSON AND BOSWORTH PROFESSOR OF ANGLO-SAXON IN THE UNIVERSITY O

OXFORD: ESS.
AT THE CLARENDO
1926.

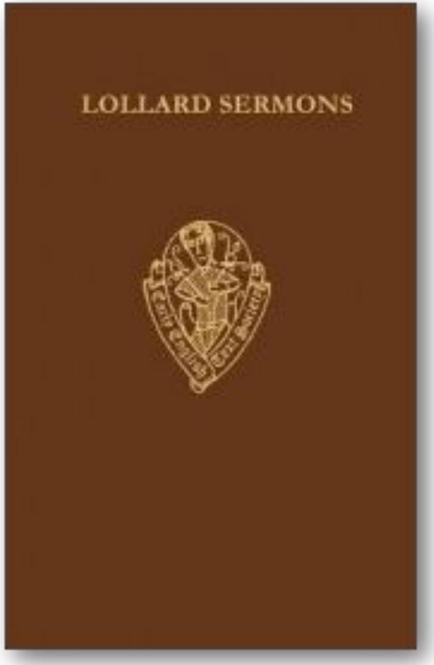
[All rights re

We’ve been here before

- They had the same problem
    - Discovered almost immediately after setting up the reading programme that the texts they were planning to use were unsuitable
        - Not available in modern editions
        - Poor or difficult-to-determine quality
    - In other words, they discovered that they needed to improve and standardise the small datasets from which they were going to draw their big data records.

We’ve been here before

- Solution? Create platform for new editions
  - Text societies and publishers to publish editions that met the NED’s requirements
  - Encouraged leading scholars to edit (and later reedit) the texts they needed
  - Very symbiotic relationship between what was going on in historical textual research at the time and the needs of this big-data dictionary
- Result was an increase in high quality small-data editions *and* better big-data data set for NED



We’ve been here before

- What we need is something similar for the digital age
  - A workflow that encourages small-data researchers to prepare their datasets in a way that
    - Respects their traditional requirements for the intensive curation and analysis of individual data points or small datasets
    - Opens these small, thick, and slow datasets up to big data analysis
    - Does not increase (and preferably reduces) the cost of production, publication, and maintenance
- **A workflow in which suitability for “big capta” research is inherent in the publication “small data” workflow rather than a separate step.**

What can I do?

- Not going to solve this problem in this paper
  - FAIR Data was the result of cross-disciplinary team working over several years through many rounds of consultation:
    - Largely focussed on a field in which data is abs/extractable
    - In which practitioners were in theory committed to data-sharing
    - In which “data” was a concept they were comfortable with
  - NED/OED was a major, society-sponsored, national effort focussed on a single data set.
- Also need to get away from idea that things would improve if only people would use my solution
  - Major issue in Humanities data is ring-fencing: my project, our club

What can I do?

- Next paper I’m going to show a conceptual prototype of one possible approach to FAIR data publication in the Humanities
  - Not *the* solution but *a* solution to something we really haven’t thought about that much.
- But real change is going to require system-wide cultural shift
  - Development of tools, systems, and practices that are themselves FAIR
  - Development of culture that considers non-FAIR data to be *infra dig*
  - But that does both of these while understanding and respecting
    - The value and purpose of Humanities research
    - The nature and purpose of Humanities data collection (including reluctance to call data “data”

Questions to ask yourself

- Are my data **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.

Questions to ask yourself

- Are my data **findable**?
- Are my data **accessible**?

A1 (meta)data are retrievable by their identifier using a standardized communications protocol [that is]

A1.1 open, free, and universally Implementable.

A1.2 allows for an authentication and authorization procedure, where necessary.

A2 metadata are accessible, even when the data are no longer available.

Questions to ask yourself

- Are my data **findable**?
- Are my data **accessible**?
- Are my data **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.

Questions to ask yourself

- Are my data **findable**?
- Are my data **accessible**?
- Are my data **interoperable**?
- Are my data **reusable**?

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.

Questions