Fostering Data Reuse: Increasing Impact and Ease in Sharing and Reusing Data

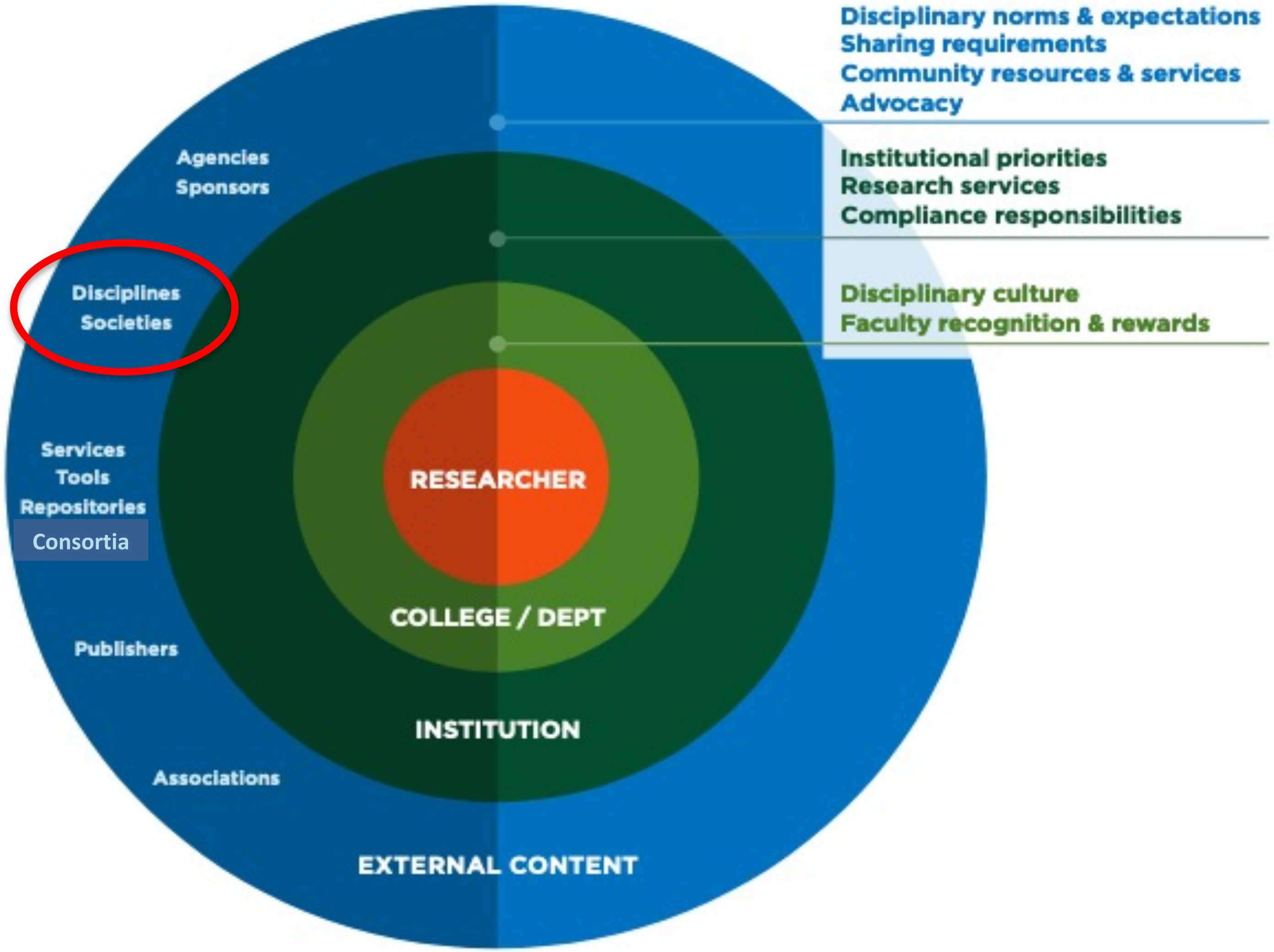
Sarah Nusser, Iowa State University and University of Virginia
Alyssa Mikytuck, University of Virginia
Gizem Korkmaz, University of Virginia

IOWA STATE UNIVERSITY

UNIVERSITY of VIRGINIA

BIOCOMPLEXITY INSTITUTE

Ecosystem influencing researcher actions



Disciplinary norms & expectations
Sharing requirements
Community resources & services
Advocacy

Institutional priorities
Research services
Compliance responsibilities

Disciplinary culture
Faculty recognition & rewards

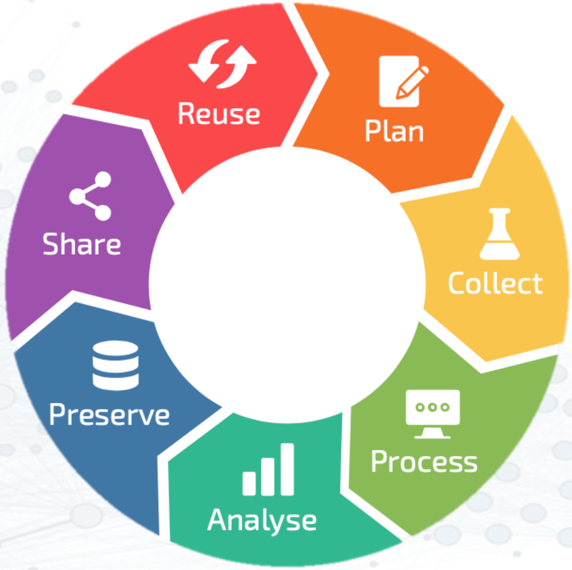
Association of American Universities and Association of Public and Land-grant Universities (2021). **Guide to Accelerate Public Access to Research Data**. Washington, DC. DOI: https://doi.org/10.31219/osf.io/tjybn CC BY- NC-SA

NSF EAGER: Practices that promote data reusability and reduce burden

National Science Foundation EAGER Award 2039677

Aims

1. What makes a data source reusable to another researcher?
2. What practices improve data reusability and burden?
3. What actions increase community readiness?



https://rdmkit.elixir-europe.org/

NSF EAGER: Practices that promote data reusability and reduce burden

National Science Foundation EAGER Award 2039677

Information gathering

Semi-structured interviews with 20 researchers

Survey of society members, ~1600 responses

- Joel Cutcher-Gershenfeld | Waymark, Brandeis

Workshop with 35 researchers and 40 stakeholders (mainly societies, funders)

Participants vary widely in:

Scholarly field

Career stage

Type of institution

Interview and Survey Participants

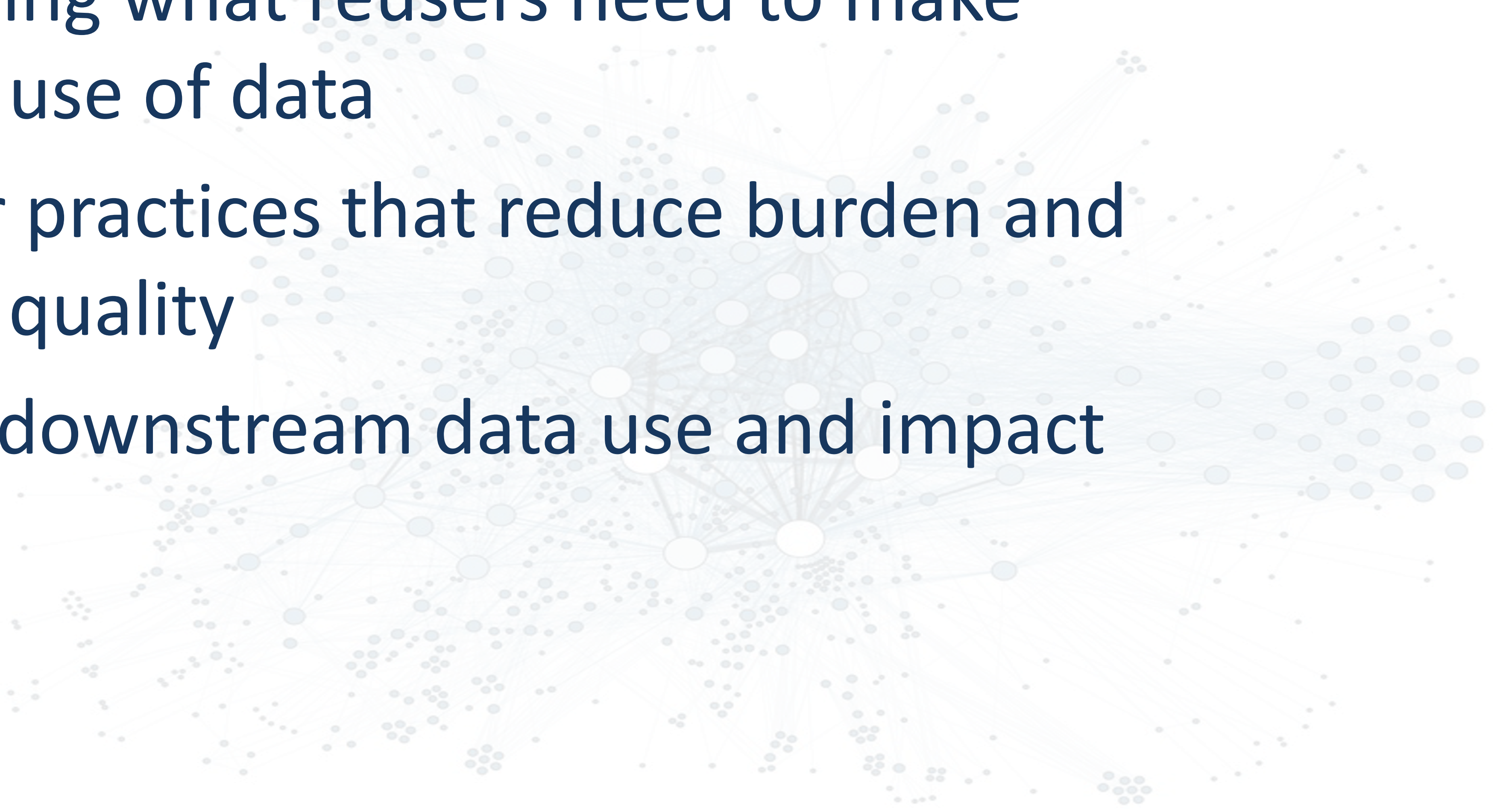
Share and reuse data ~60%

Reuse only ~30%

Produce only ~10%

Focal Areas

1. Anticipating what reusers need to make effective use of data
2. Producer practices that reduce burden and facilitate quality
3. Tracking downstream data use and impact



1. Anticipating what reusers need to make effective use of data

How important / easy is

ensuring clear, accurate, and complete data documentation are associated with the shared data?

Important 9.0

Easy 2.8

(0-10 scale)

User Experiences

Lack of complete information

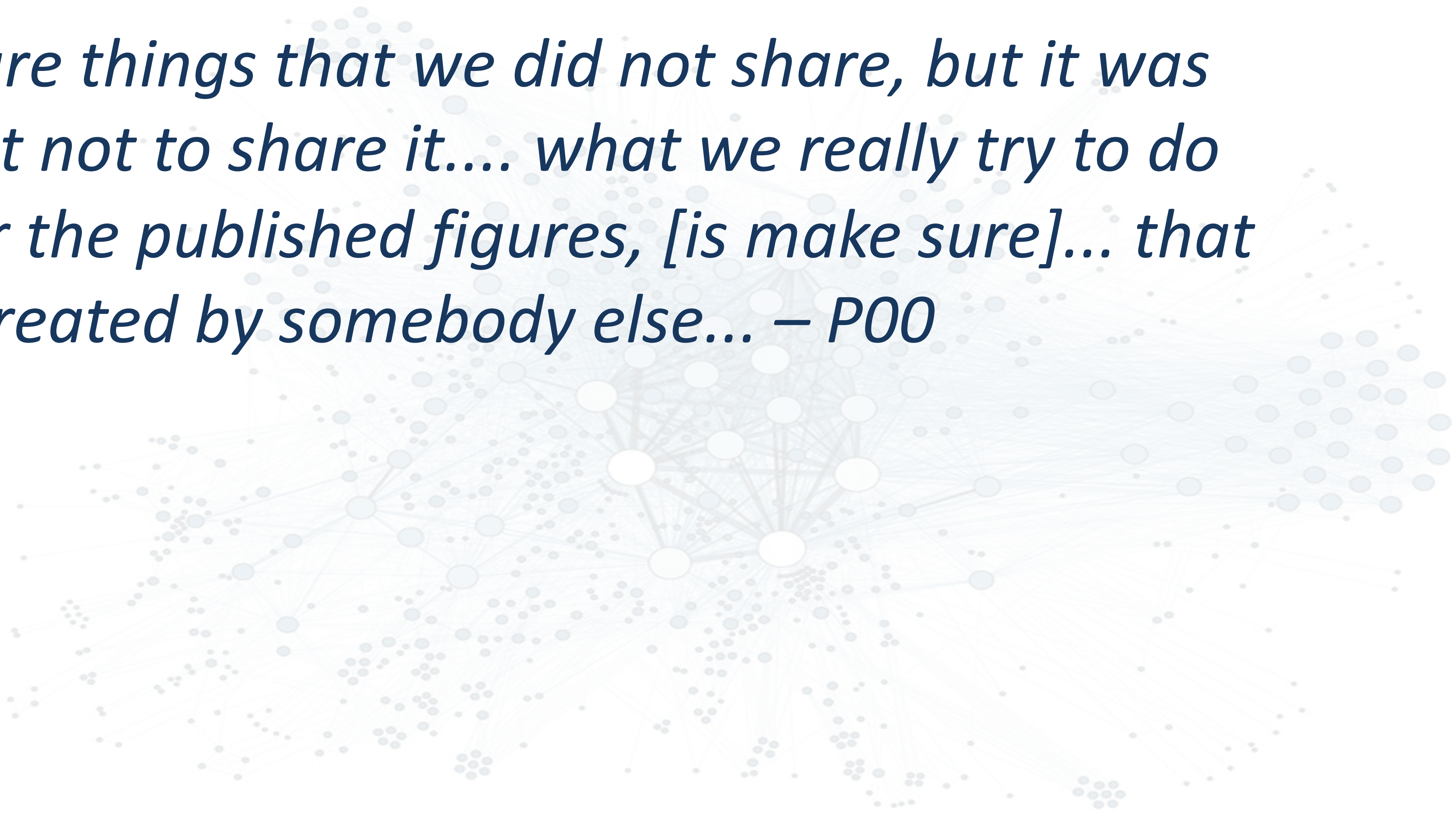
Sometimes we find papers mentioning publicly available research datasets, and then we find the data is not annotated, and that annotation is not available... which means the data itself is useless for our purpose – B04

Producers may be unresponsive to data or clarification requests

Producer Uncertainties

What to share with data

I'm sure there are things that we did not share, but it was not out of intent not to share it.... what we really try to do is, especially for the published figures, [is make sure]... that they can be recreated by somebody else... – P00



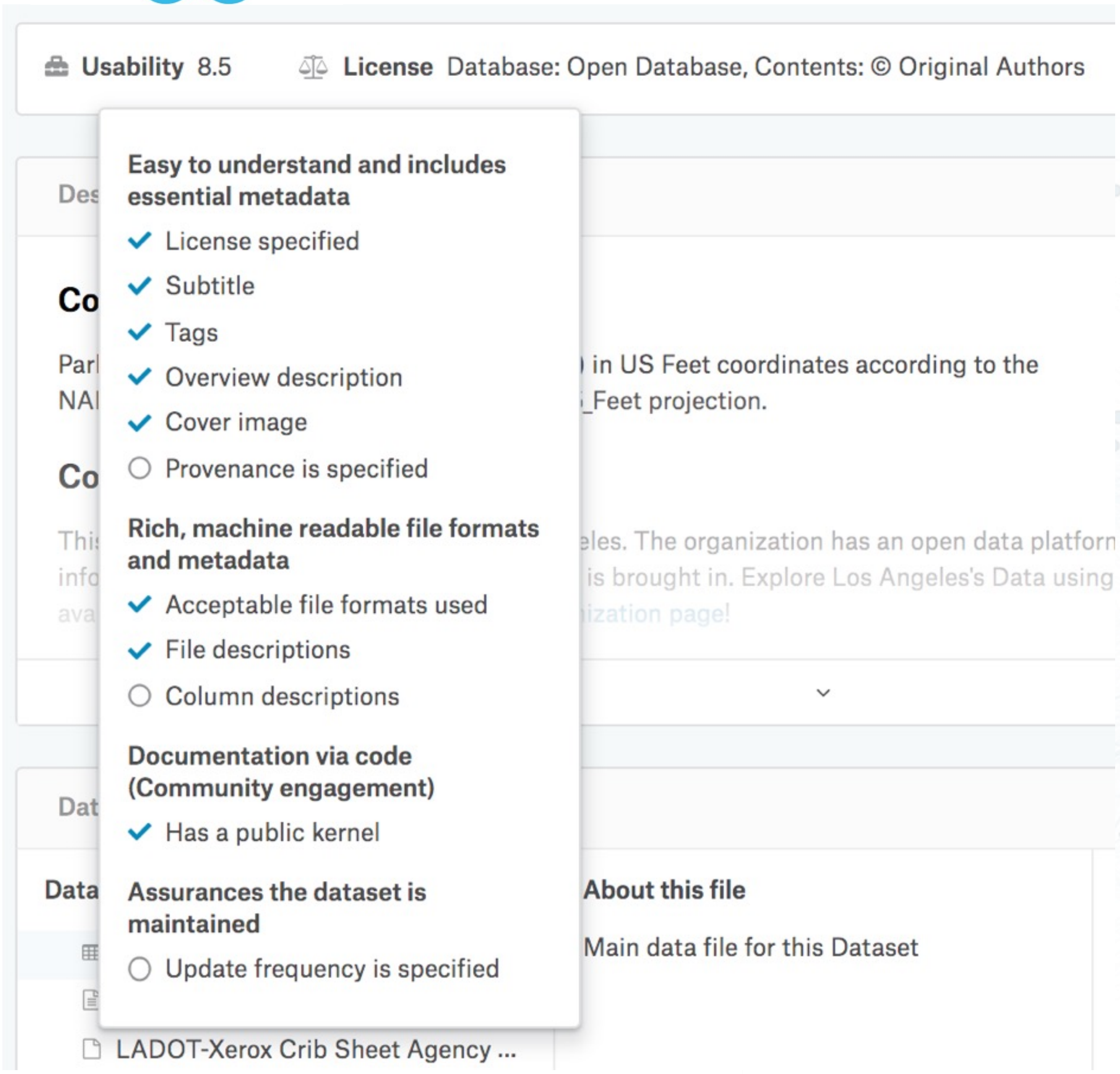
Context needed by reusers

Faniel, *et al*. (2019), Faniel & Yakel (2017)

Data reuser assessments	**Relevant context**
Relevance to study objectives	**Study documentation**: study objectives, methodology report, publication
Trustworthy, credible	**Reputations**: producer, institution, repository, others' use of the shared data source
Understand data contents	**Data documentation**: provenance, metadata, standards used + study information (above)
Understand how data were generated	**Methods information**: methodology report, code for acquiring and preparing data for analysis
Understand permissible uses	**Use specifications and restrictions**: data license/use agreement, possibly IRB
Evaluate data quality and potential issues	**Issues/remedies documentation**: study documentation, data documentation, code
Understand how to analyze data	**Data use information**: study documentation, code
Ability to provide cite data and credit producer	**Citation/credit information**: citation reference, persistent identifiers [e.g., data (DOI), producers (ORCID), institutions (ROR), funders (Crossref), …]

Data Science for the Public Good Repository Project

kaggle



Usability 8.5   License Database: Open Database, Contents: © Original Authors

Des

Co

Parl) in US Feet coordinates according to the
NAI _Feet projection.

Co

This eles. The organization has an open data platform
info is brought in. Explore Los Angeles's Data using
ava ization page!

Easy to understand and includes essential metadata

- ✓ License specified
- ✓ Subtitle
- ✓ Tags
- ✓ Overview description
- ✓ Cover image
- ○ Provenance is specified

Rich, machine readable file formats and metadata

- ✓ Acceptable file formats used
- ✓ File descriptions
- ○ Column descriptions

Documentation via code (Community engagement)

- ✓ Has a public kernel

Assurances the dataset is maintained

- ○ Update frequency is specified

Dat

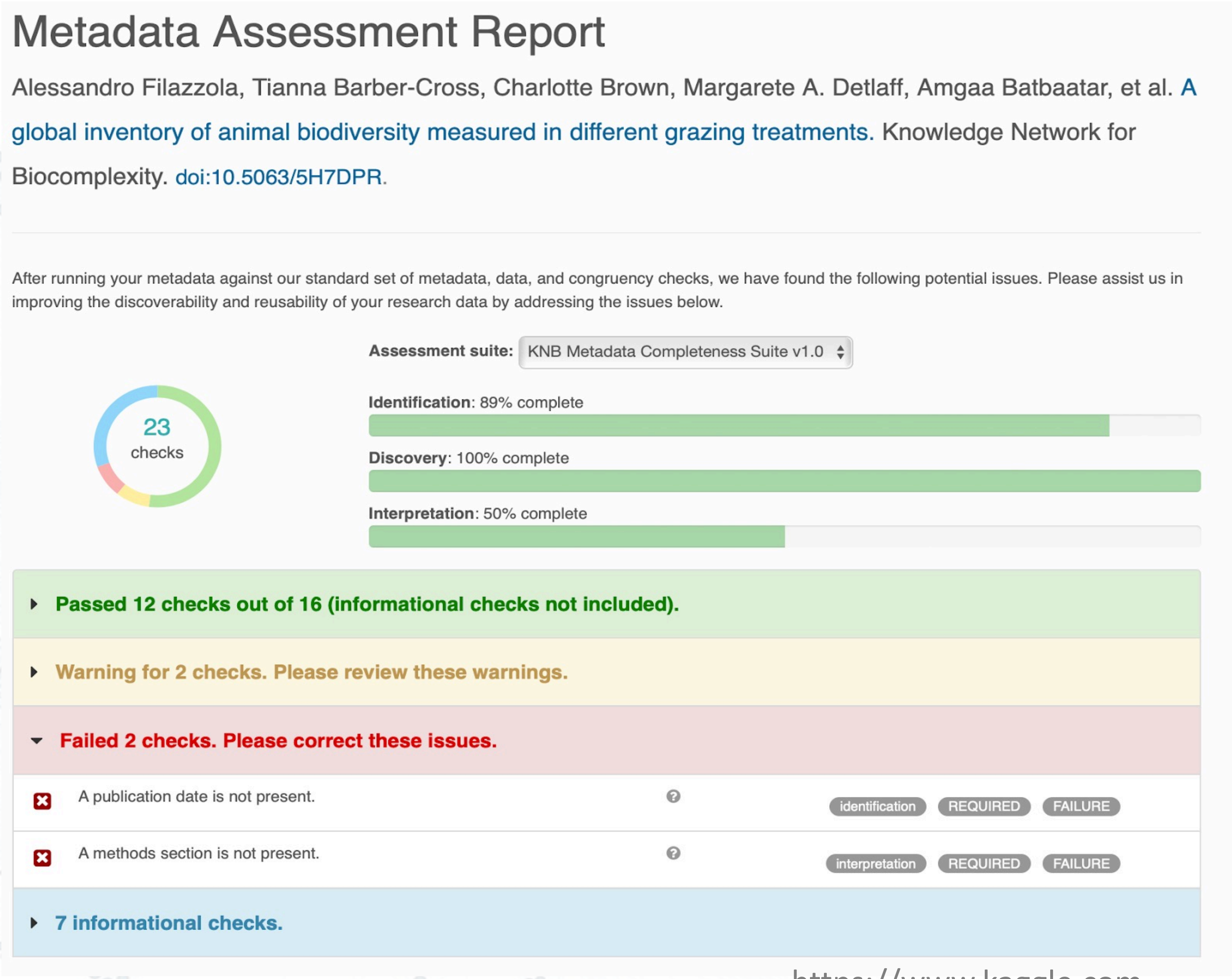
Data

About this file

Main data file for this Dataset

LADOT-Xerox Crib Sheet Agency ...

knb



Metadata Assessment Report

Alessandro Filazzola, Tianna Barber-Cross, Charlotte Brown, Margarete A. Detlaff, Amgaa Batbaatar, et al. A global inventory of animal biodiversity measured in different grazing treatments. Knowledge Network for Biocomplexity. doi:10.5063/5H7DPR.

After running your metadata against our standard set of metadata, data, and congruency checks, we have found the following potential issues. Please assist us in improving the discoverability and reusability of your research data by addressing the issues below.

Assessment suite: KNB Metadata Completeness Suite v1.0

Identification: 89% complete

Discovery: 100% complete

Interpretation: 50% complete

▸ **Passed 12 checks out of 16 (informational checks not included).**

▸ **Warning for 2 checks. Please review these warnings.**

▾ **Failed 2 checks. Please correct these issues.**

☒	A publication date is not present.	identification	REQUIRED	FAILURE
☒	A methods section is not present.	interpretation	REQUIRED	FAILURE

▸ **7 informational checks.**

https://www.kaggle.com
https://knb.ecoinformatics.org

1. Anticipating reuser needs for using shared data

Practice recommendation

Establish a standard set of expected context to be shared with data

Societies

Spearhead definition of field-specific data documentation expectations

Journals, repositories

Require standard documentation set with shared data source

Repositories

Assess the potential usability of the data source

2. Establishing producer practices that reduce burden and facilitate quality

Producer Experiences

Lack of planning and its impact

When did you start taking steps that would facilitate sharing it?

Not nearly soon enough *… it's definitely at the backend, but that's probably one of the reasons that we only shared the final analysis data set. – B10*

Importance of starting early

*[Getting] a **team of researchers that agree on how** the data's going to be collected, what the quality standards are, ... That's all got to be done in advance **before the data is collected**. – P01*

*I think that you really have to **start before the data is gathered if you want to do a good job of making it available** – B09*

How important / easy is

planning for how data will be documented and shared before you start your research study?

Important 6.7

Easy 4.1

(0-10 scale)

Planning and Process Strategies

- Focus on own future reuse – *P00*
- Implement data sharing for own research group – *B03*
- Standardize and automate – *P00, B05, B09, others*
  - Plan detailed workflow to standardize process
  - Adopt or create standards for data elements, formats
  - Develop pipelines to automate data intake and checking

There is currently **sufficient knowledge and training** in my primary field or discipline on **software and tools that can reduce burden in producing and documenting research data**

23% had high agreement (7-10)

(0-10 scale)

Potential paths forward

- Establish standard data preparation/sharing process as the basis for planning
- Develop training on data preparation/workflow approaches, tools
- Perspectives on how
  - Training: undergrad/graduate education, for all career stages *(B05)*
  - Team science approach: research data expert with disciplinary and data science knowledge as primary team member
  - Research data services (EU: Data Stewardship Competency Centers)

2. Establishing practices to reduce burden, facilitate quality

Practice recommendations

Adopt **early planning for the how** of documenting data for reuse and promoting data quality

Develop an **automated workflow mindset**: define research process as standardized and automated approach to collecting/checking data and documentation

Societies, Funders

Important convening opportunities to discuss

- What elements should be in a planning process to leverage automated workflows for data/doc capture and preparation
- Realistic, sustainable approaches to adopting planning/workflows

3. Tracking downstream data use and impact

The tenure, promotion, and rewards in my organization recognize and value researchers for sharing research data

26% Very Strongly Disagreed (0)

(0-10 scale)

How reusers reference data used

Text-based approaches to citing data

The original paper was something that could be easily referenced, just like any other publication. And we added a phrase [in the text] that we used this particular data from the accompanying data of this publication. – U04

[In the publication text], I listed in the methods, described the data source and the citation for the survey protocol document, as well as listed the website where the data could be accessed. – U03

Producers tracking reuse of shared data

Lack awareness of shared data reuses

Do you have any sense of what kinds of uses are occurring as a result of your shared data or how you would find that out?

No, I don't have a good idea of that. *The only way I think I would know that [is] if someone reached out to me, "Hey, I downloaded the data and used it. I have a question or things."* ***[There’s] nothing at the*** **system** ***level that tells me how they're being used***. *– B10*

Academic credit for a researcher

Enormous bias in favor of **publications** (authorship, journals)

- What would drive first-class status for shared data?

Outputs assessed for researcher’s **contribution, quality of shared output,** and its **impact**

- How are these attributes assessed for data?
- Where would that information come from?

3. Tracking downstream data use and impact

Practice recommendation

Reusers: Cite data and producers

Producers: Select repositories that collect information to evaluate impact
Document shared data and its future use in CV

Societies

Contribute to defining how data contribution, quality, impact is assessed

Journals, repositories

Establish expectations and standard paths for citation and credit

Monitor and share downstream actions of shared data

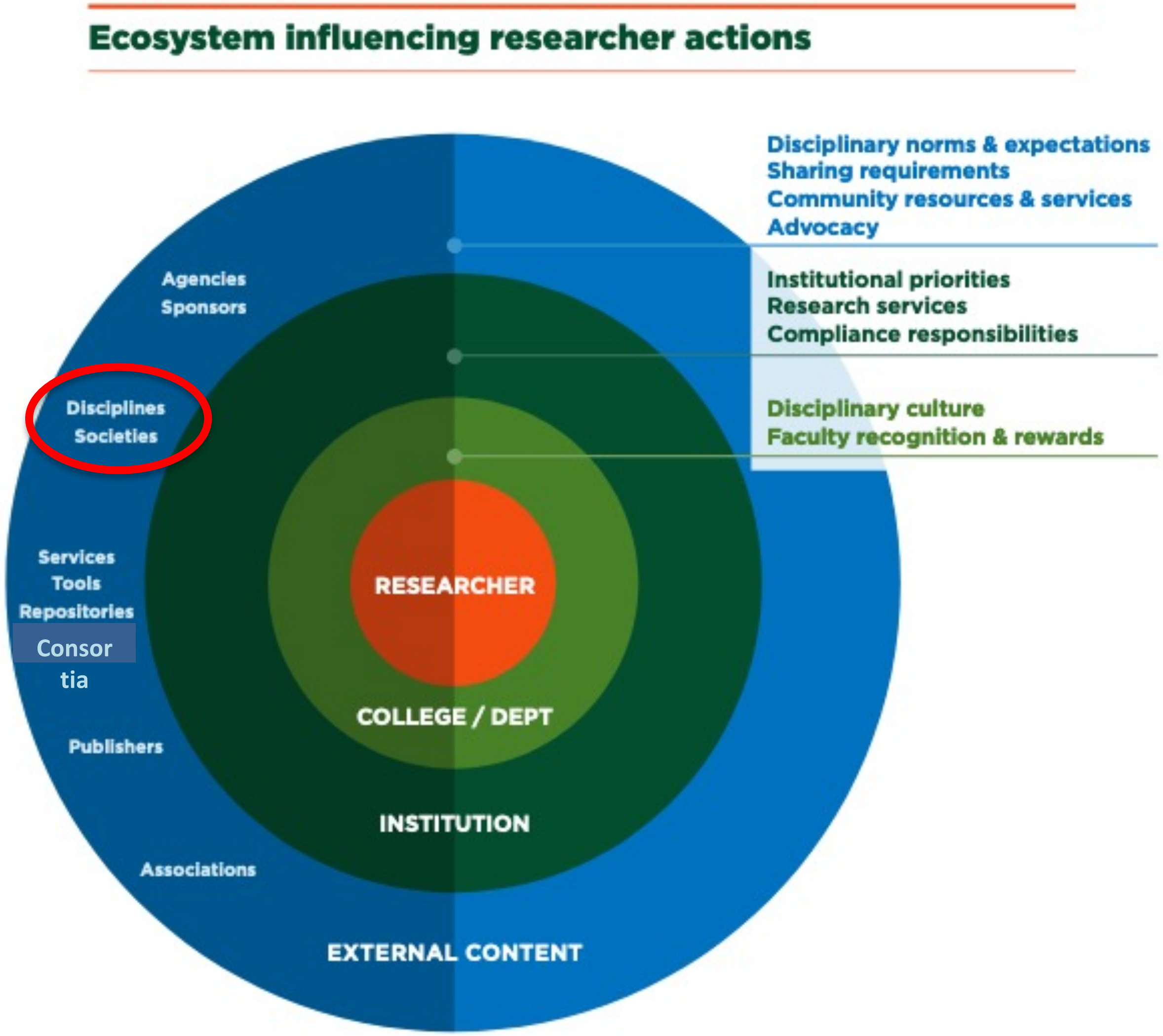
Funders

Value shared data and impact metrics in proposals and progress reports

Path Forward

What is the vision we aspire to?

What role can societies play in increasing impact and ease in data sharing?



Association of American Universities and Association of Public and Land-grant Universities (2021). **Guide to Accelerate Public Access to Research Data**. Washington, DC. DOI: https://doi.org/10.31219/osf.io/tjybn CC BY- NC-SA

Shared Vision for Success

- Data sharing is rigorously executed and expected part of the research process across fields and disciplines
- Infrastructure for automation of metdata generation [and data capture/checking] is built into the digital tools used in the research process
- When you read an article that uses a data set, a DOI links to a data source and the code to use the data in real-time
- The creation and curation of data is valued as much as the clever analysis of data

Society Actions

Expected data sharing standards

- Define contextual documentation, formats, citation information, repository features, … required for shared data

Journal data sharing policies

- Develop and require community-based discipline-specific practices in data sharing, building across journals in a field

Culture and recognition / data as first-class product

- Clarify what kind of credit researchers want & meaningful metrics
- Develop recommended approach for documenting and evaluating shared data in promotion (other reward processes)

Thank you!

Questions?

Sarah Nusser, Iowa State University and University of Virginia
Alyssa Mikytuck, University of Virginia
Gizem Korkmaz, University of Virginia

NSF

IOWA STATE UNIVERSITY

UNIVERSITY of VIRGINIA

BIOCOMPLEXITY INSTITUTE