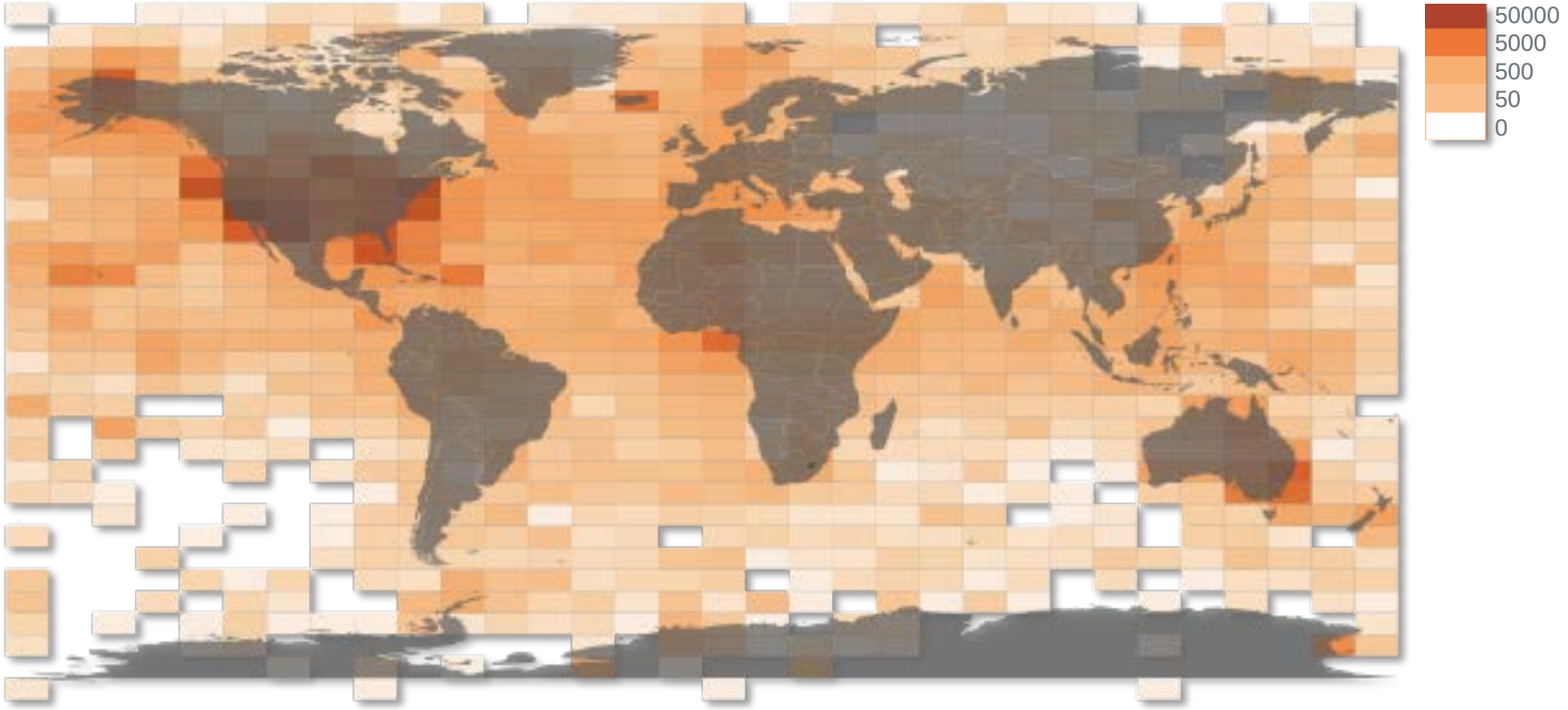
QUANTIFYING FAIR: AUTOMATED METADATA IMPROVEMENT AND GUIDANCE IN THE DATAONE REPOSITORY NETWORK

Matthew B. Jones, Peter Slaughter, Ted Habermann

0000-0003-0077-4738
@metamattj

Implementing FAIR Data for People and Machines: Impacts and Implications, Sep 11, 2019

DataONE



Global Data Coverage

DataONE

Repository Federation

Global

Interoperable

Community

DataONE Metrics

Global
Data Coverage

806K
Data Packages

40 Webinars

10
Education Modules

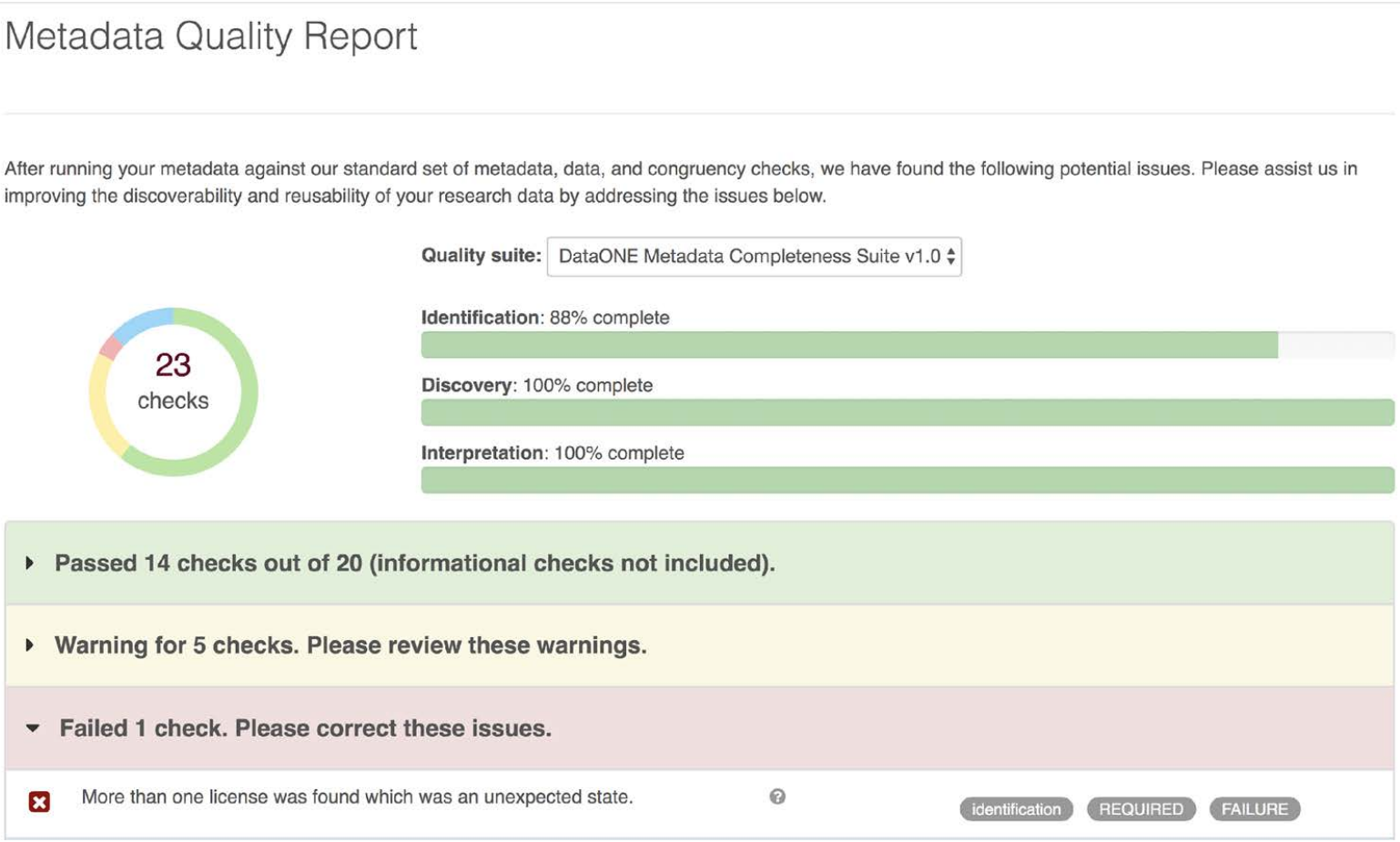
42
Member Repos

152K
Contributors

19,600
Users/Month

5300+
Trained

MetaDIG: Metadata Improvement and Guidance



Findable

Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services.

Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

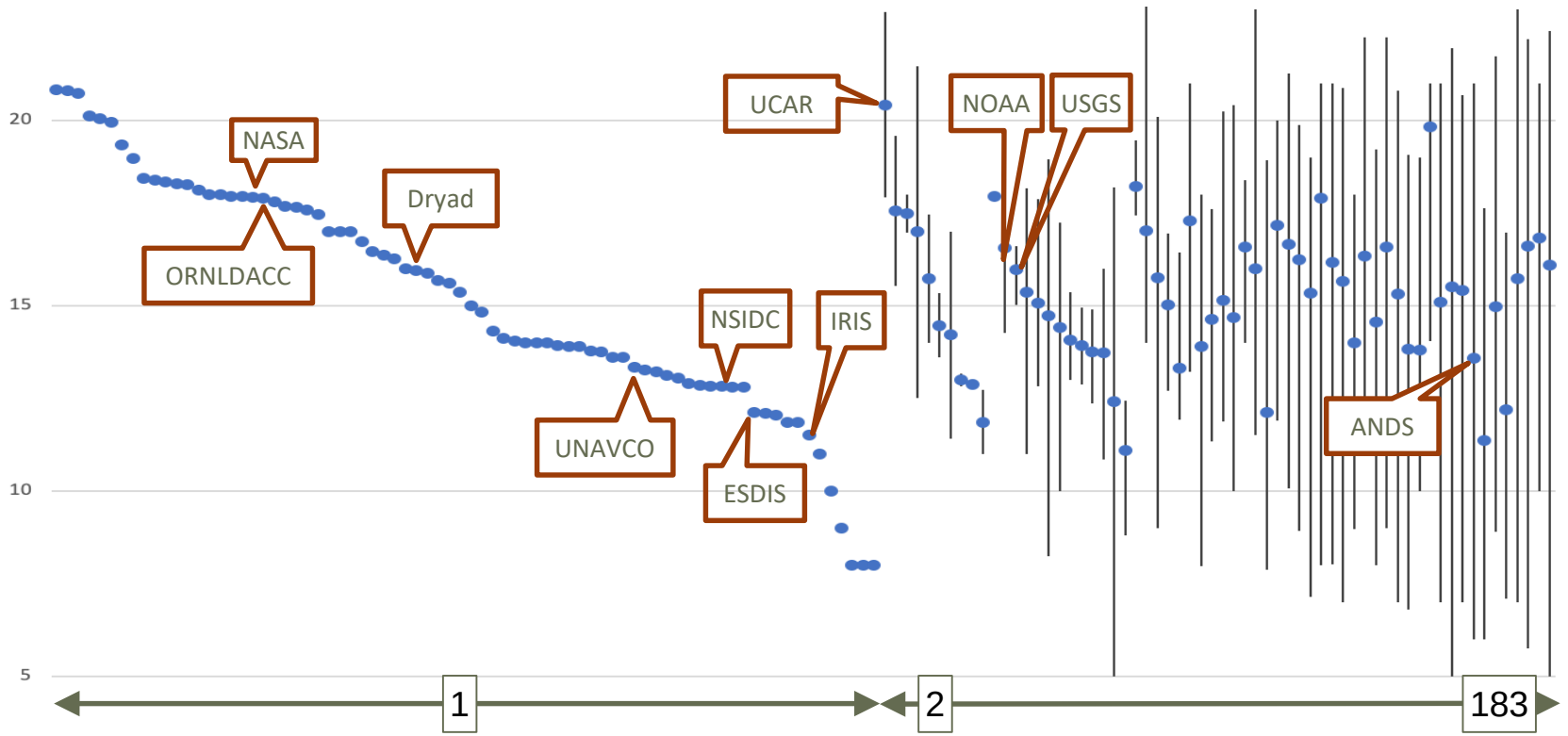
Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

Reusable

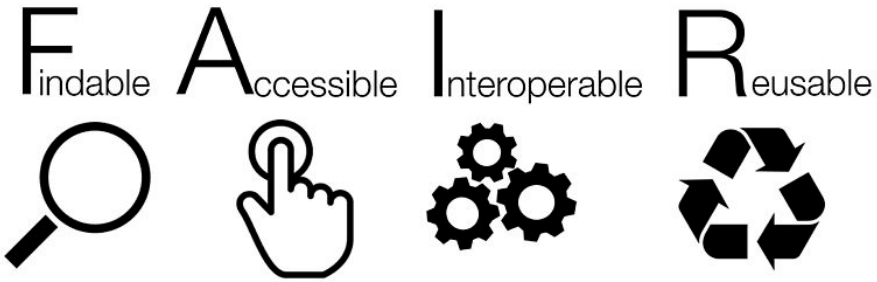
The ultimate goal of FAIR is to optimise the reuse of data. Metadata and data should be well-described so they can be replicated and combined in different settings.

FAIR Measure – DataCite Providers



Slide credit: Ted Habermann

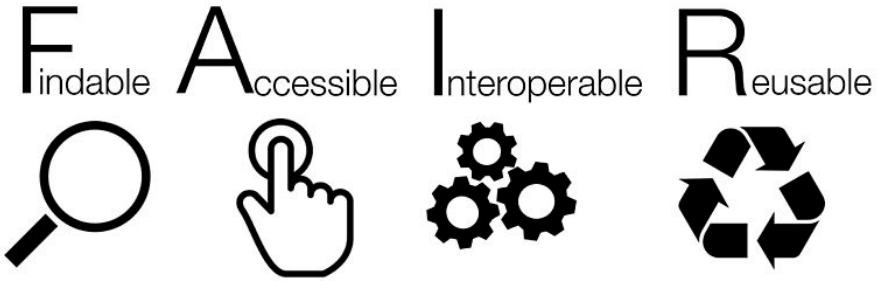
Number of Clients



“A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a **concise and measurable** set of principles that we refer to as the FAIR Data Principles.” Wilkinson et al., 2016

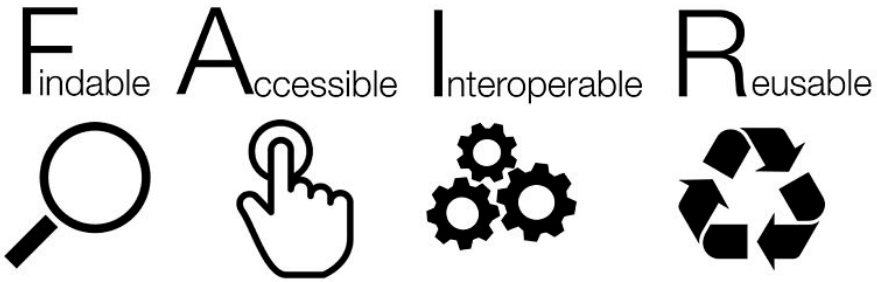
F2. Data are described with **rich metadata** (defined by R1 below)

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



Binary?

Yes or No?
True or False?



Continuum

Findable

Accessible

Interoperable

Reusable

FAIR metrics, a community process

Wilkinson et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018. https://doi.org/10.1038/sdata.2016.18

FAIR Metrics workshop March 2019

- Deep dive into metadata concepts
  - Ecological Metadata Language
  - ISO 19115*
  - DataCite metadata

FAIR Metrics session, July 2019

- Community consensus via Documentation cluster
- Discussed > 90 FAIR checks
- Implemented 52 checks

- https://github.com/NCEAS/metadig-checks/

Findable

DataONE FAIR Checks

17 Checks

Item that is checked	Description of check	Facet	Required	Implemented
title	presence, length, content	F2	Y	partially
metadata identifier	presence, globally unique, id type	F1	Y	partially
resource identifier	presence, globally unique, id type	F3	Y	partially
resource identifier type	presence	F3	Y	Y
publication date	presence	F2	Y	Y
abstract	presence, length, content	F2	Y	partially
award # or funder	presence	F2	N	Y
temporal coverage	presence	F2	N	Y

Accessible

DataONE FAIR Checks

10 Checks

Item that is checked	Description of check	Facet	Required	Implemented
publisher	presence, significant name, is it an organization id?	A1	Y	partially
distributor	presence, significant name, is it an organization id?	A1	Y	partially
identifier	retrievable	A1	Y	N
resource distribution URL for landing page	presence, retrievable, protocol type	A1	Y	partially
service data url	presence, retrievable, protocol type	A1	Y	N

Interoperable

DataONE FAIR Checks

12 Checks

Item that is checked	Description of check	Facet	Required	Implemented
metadata schema	the metadata document is schema valid	I1	Y	N
data format	presence, data in non-proprietary format	I1	Y	partially
checksum	presence, checksum matches data		Y	partially
attribute definition	presence	I2	Y	Y
attribute names unique	for an entity, names are unique	I2	Y	N
attribute storage type	presence	I2	Y	Y

Reusable

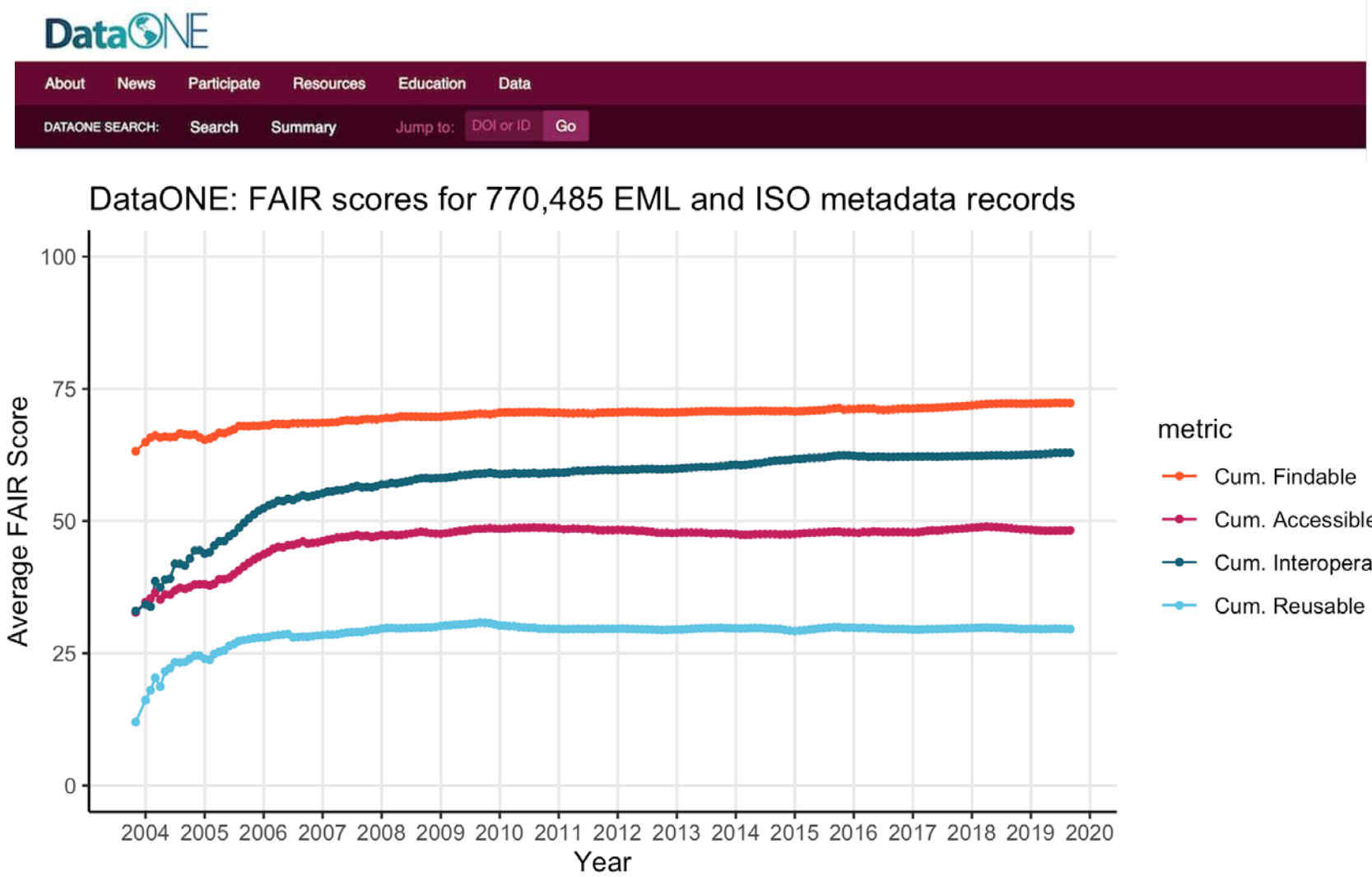
DataONE FAIR Checks

13 Checks

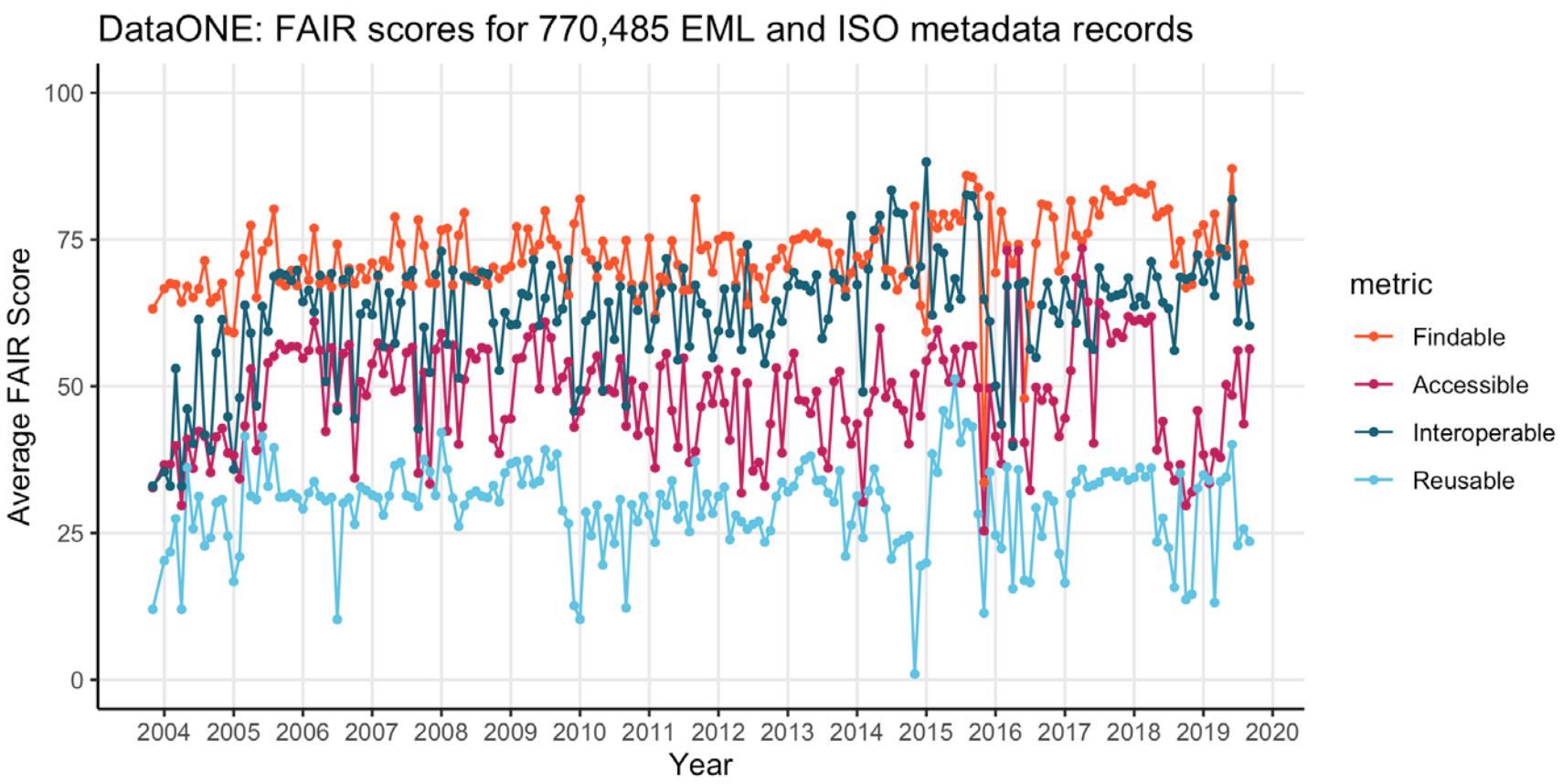
Item that is checked	Description of check	Facet	Required	Implemented
metadata license	presence	R1.1	Y	Y
data license	presence	R1.1	Y	Y
resource description	presence		Y	Y
methods description	presence		Y	Y
attribute units	presence, controlled vocabulary	R1.3	Y	partially
attribute domain	presence, congruence	R1.3	Y	partially
attribute measurement scale	presence	R1.3	Y	Y

Are datasets in DataONE FAIR?
Preliminary results

Wilkinson et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018. https://doi.org/10.1038/sdata.2016.18


DataONE
About
News
Participate
Resources
Education
Data
DATAONE SEARCH:
Search
Summary
Jump to:
DOI or ID
Go
DataONE: FAIR scores for 770,485 EML and ISO metadata records
Average FAIR Score
100
75
50
25
0
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year
metric
Cum. Findable
Cum. Accessible
Cum. Interoperable
Cum. Reusable

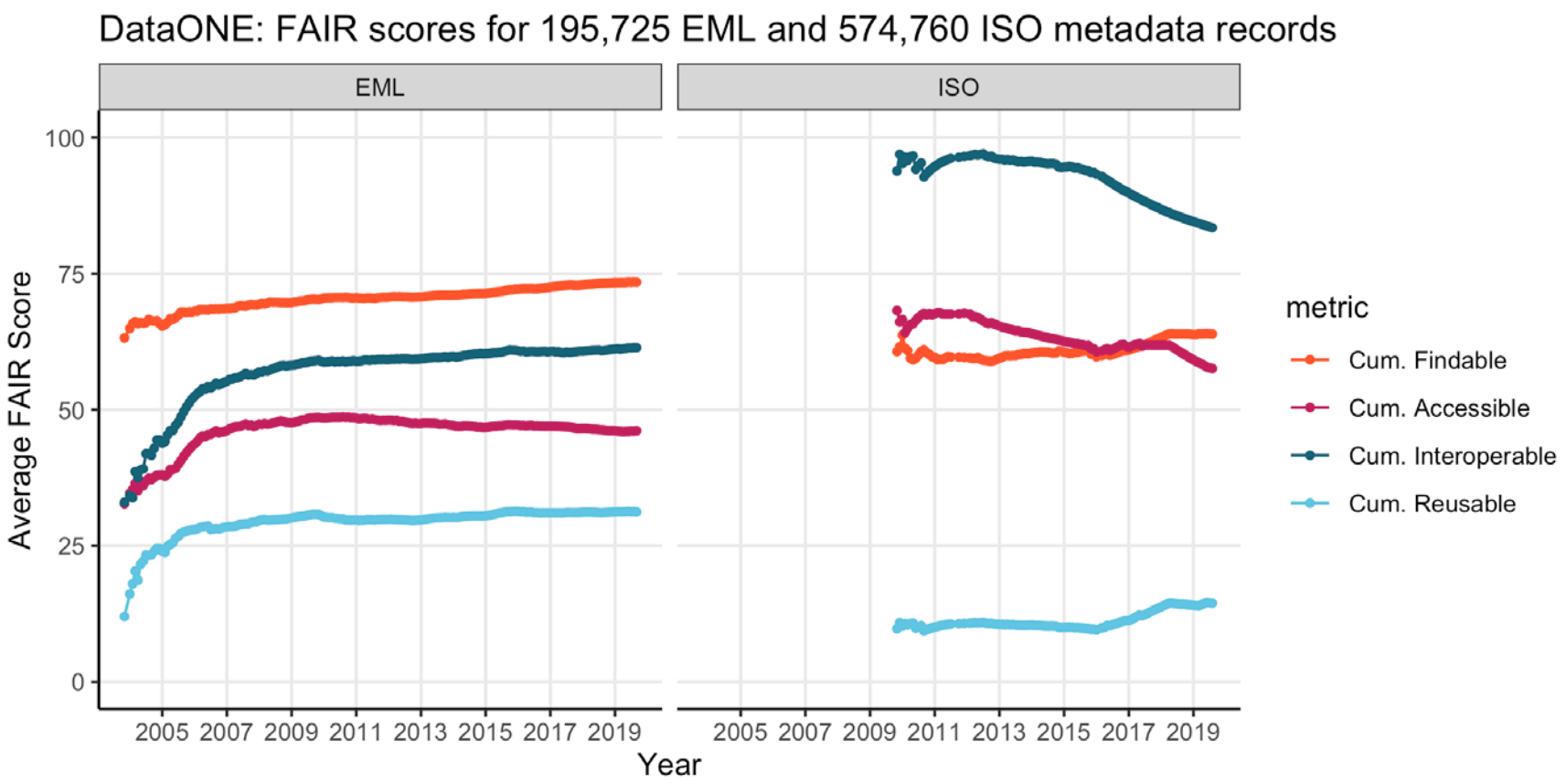

DataONE
About
News
Participate
Resources
Education
Data
DATAONE SEARCH:
Search
Summary
Jump to:
DOI or ID
Go

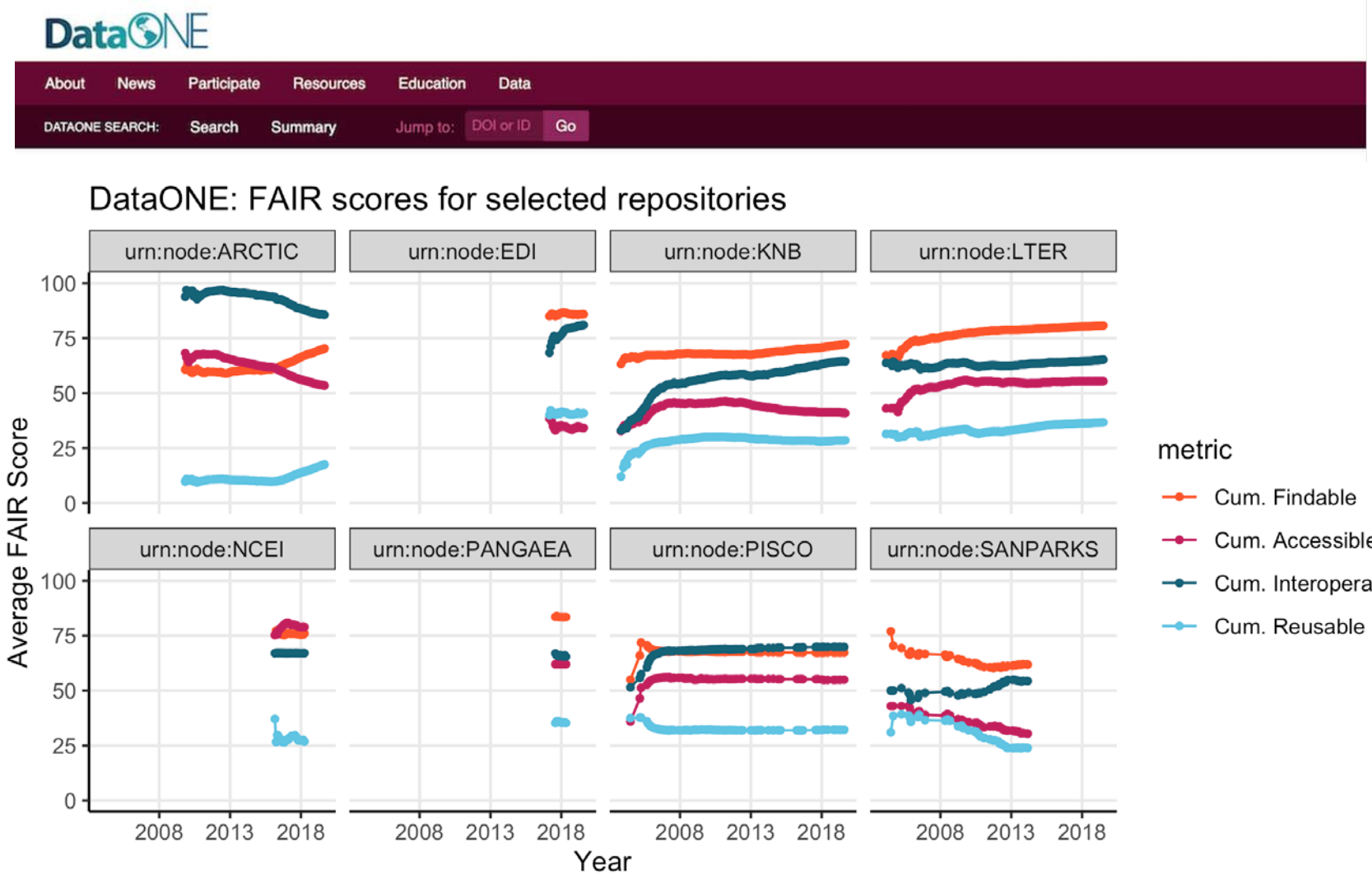

DataONE: FAIR scores for 770,485 EML and ISO metadata records
Average FAIR Score
100
75
50
25
0
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year
metric
Findable
Accessible
Interoperable
Reusable

DataONE

About News Participate Resources Education Data

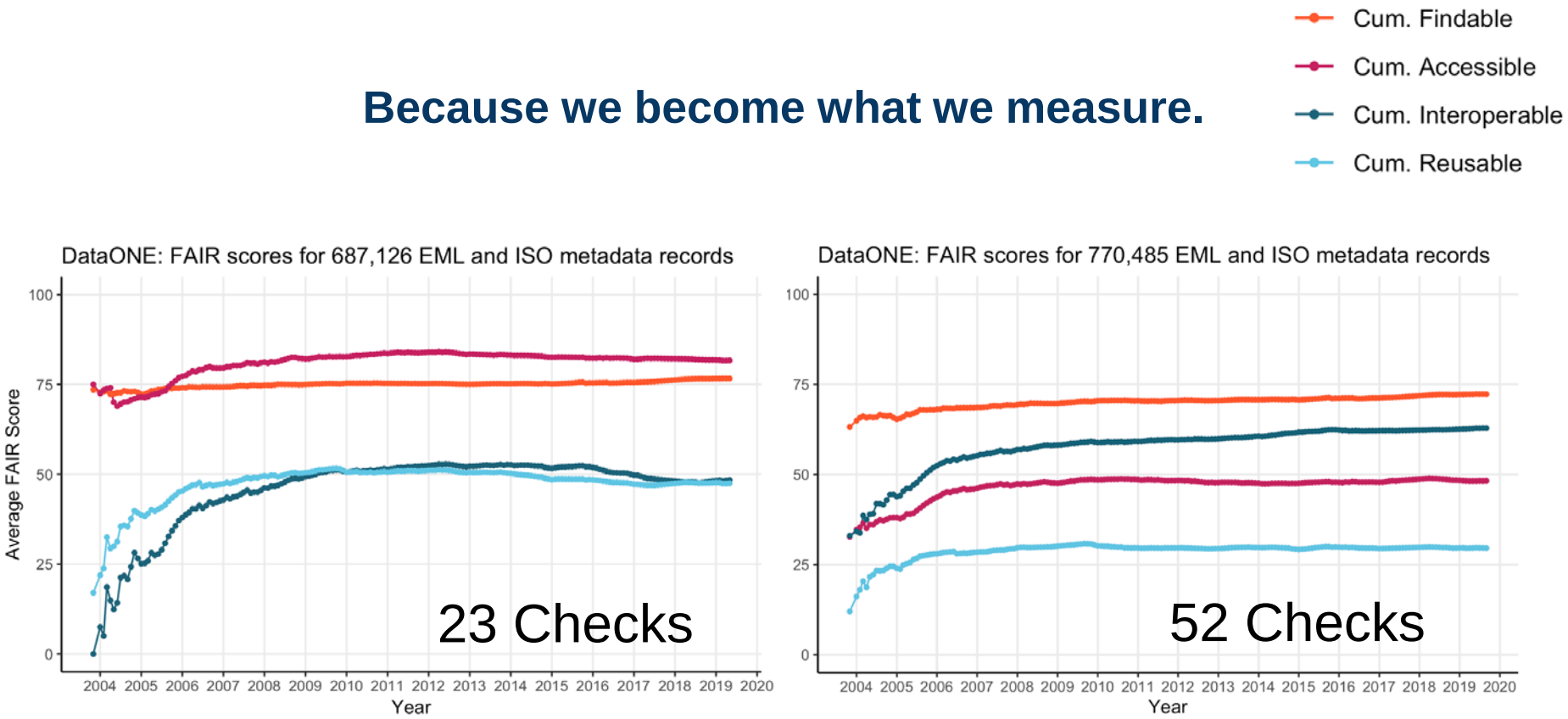
DATAONE SEARCH: Search Summary Jump to: DOI or ID Go




DataONE
About
News
Participate
Resources
Education
Data
DATAONE SEARCH:
Search
Summary
Jump to:
DOI or ID
Go
DataONE: FAIR scores for selected repositories
urn:node:ARCTIC
urn:node:EDI
urn:node:KNB
urn:node:LTER
urn:node:NCEI
urn:node:PANGAEA
urn:node:PISCO
urn:node:SANPARKS
Average FAIR Score
100
75
50
25
0
2008
2013
2018
Year
metric
Cum. Findable
Cum. Accessible
Cum. Interoperable
Cum. Reusable

Why Community Consensus?

Because we become what we measure.



www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

Comment: A design framework and exemplar metrics for FAIRness

Mark D. Wilkinson[1], Susanna-Assunta Sansone[2], Erik Schultes[3], Peter Doorn[4], Luiz Olavo Bonino da Silva Santos[5,6] & Michel Dumontier[7]

- Clear
- Realistic
- Discriminating
- Measurable
- Universal

https://doi.org/10.1038/sdata.2018.118

Modeling the FAIR Rubrics Landscape

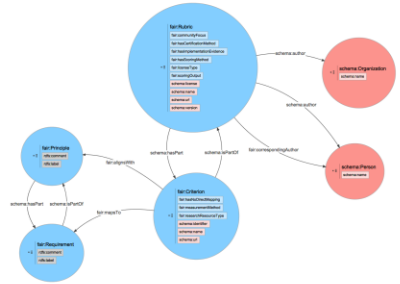
https://osf.io/685sw/

RDAP Summit
2019.05.15-2019.05.17
Coral Gables, FL

Marijane White[1], Lily Winfree[2], Payal Mehndiratta[3], Kimberly Robasky[3,4,5], Robin Champieux[1]

[1]Library, Oregon Health & Science University, Portland, OR; [2]Open Knowledge International; [3]Rennaissance Computing Institute, [4]Department of Genetics, [5]School of Library and Information Science, University of North Carolina, Chapel Hill, NC

Figure 1: Semantic Data Model



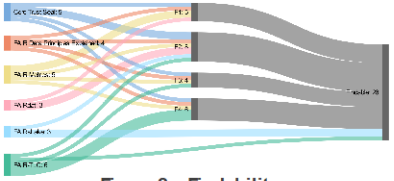


Figure 2a: Findability

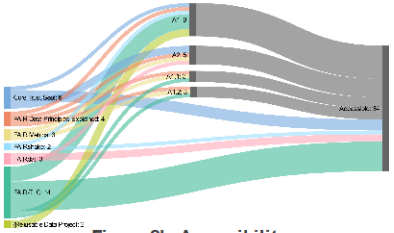


Figure 2b: Accessibility

What

The FAIR Data Principles[1] are the gold standard for evaluating the management and sharing of data and research resources. Many parallel efforts have emerged to identify recommended practices and metrics to help researchers and institutions improve and measure the FAIRness of their sharing efforts.

In this work, we conducted an exploratory evaluation of seven rubrics that interpret the FAIR Data Principles and how to meet them:

a) Core Trust Seal[2]
b) FAIR Data Principles Explained[3]
c) FAIR Metrics[4]
d) FAIRdat[5]
e) FAIRshake[6]
f) FAIR-TLC[7]
g) (Re)usable Data Project[8]

Collectively, the rubrics have 167 criteria that either align with the Principles or map directly to their requirements. Some criteria align with or map to more than one Principle or requirement, and nine criteria do not align with or map to any of them.

Why

The FAIR principles are good but they can be difficult to interpret. The principles themselves do not articulate specific practices or actions, but there is a growing body of rubrics that give specific recommendations and guidelines for adhering to the principles. We wanted to understand and help people act upon the different ways

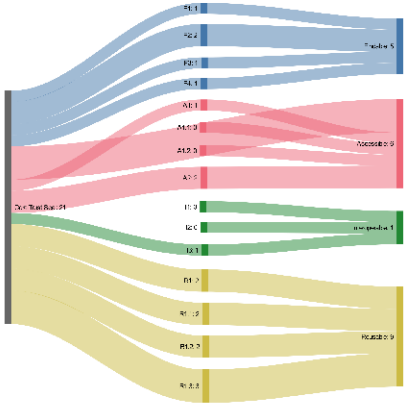


Figure 3a: Core Trust Seal

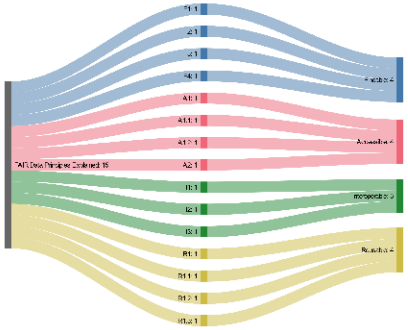


Figure 3b: FAIR Data Principles Explained

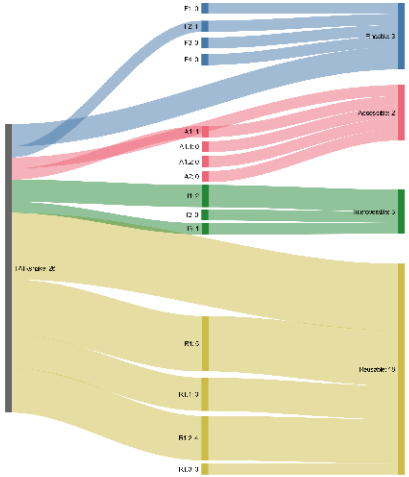
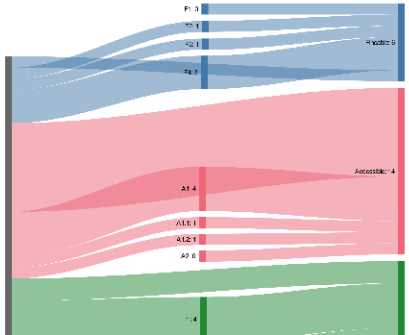


Figure 3e: FAIRshake



FAIR is concise

FAIR is ambiguously measurable

FAIR is a continuum

We need community consensus

We will become what we measure

Big thanks to our collaborators:

Ted Habermann
Sean Gordon
Margaret O'Brien
Bryce Mecum
Amber Budden
Dave Vieglais
and
the DataONE Team

MetaDIG: 1443062
Arctic Data Center: 1546024
DataONE: 1421275

DataONE