Quantifying FAIR: metadata improvement and guidance in the DataONE repository network

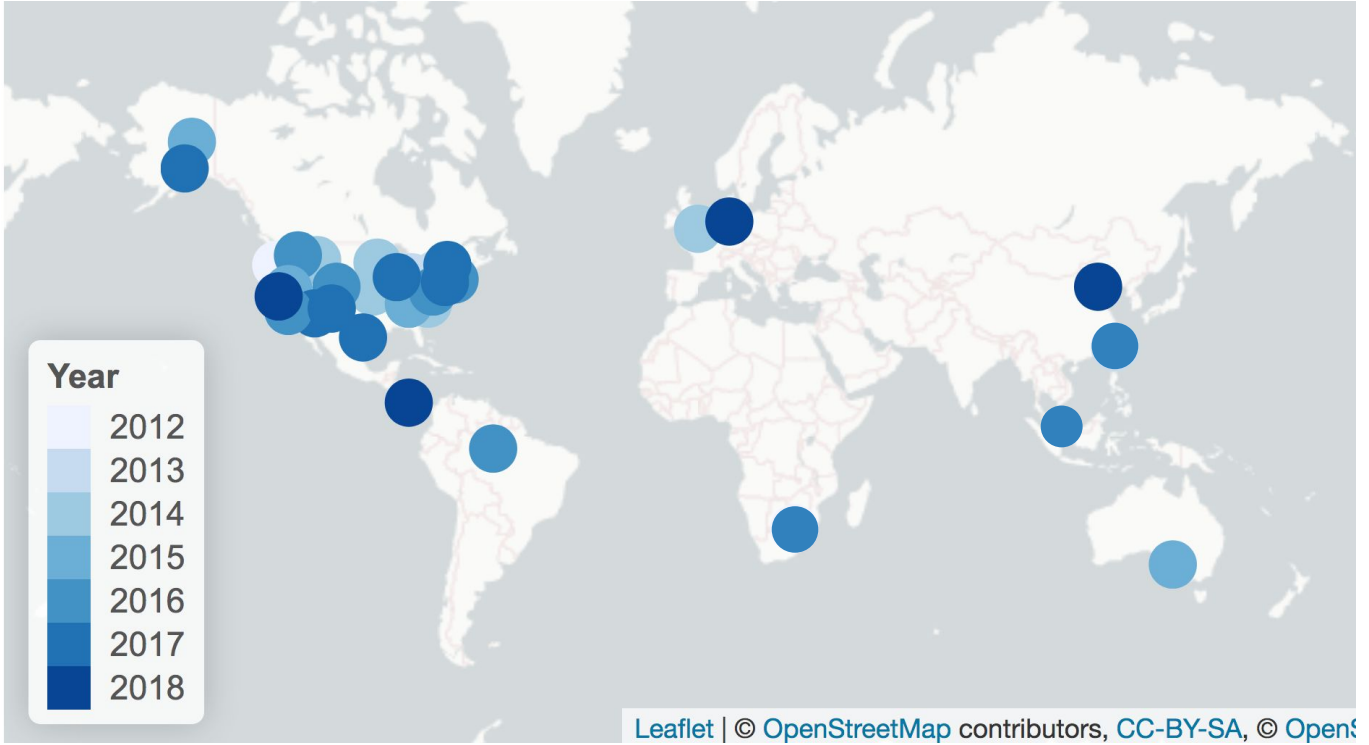
Matthew B. Jones
Peter Slaughter

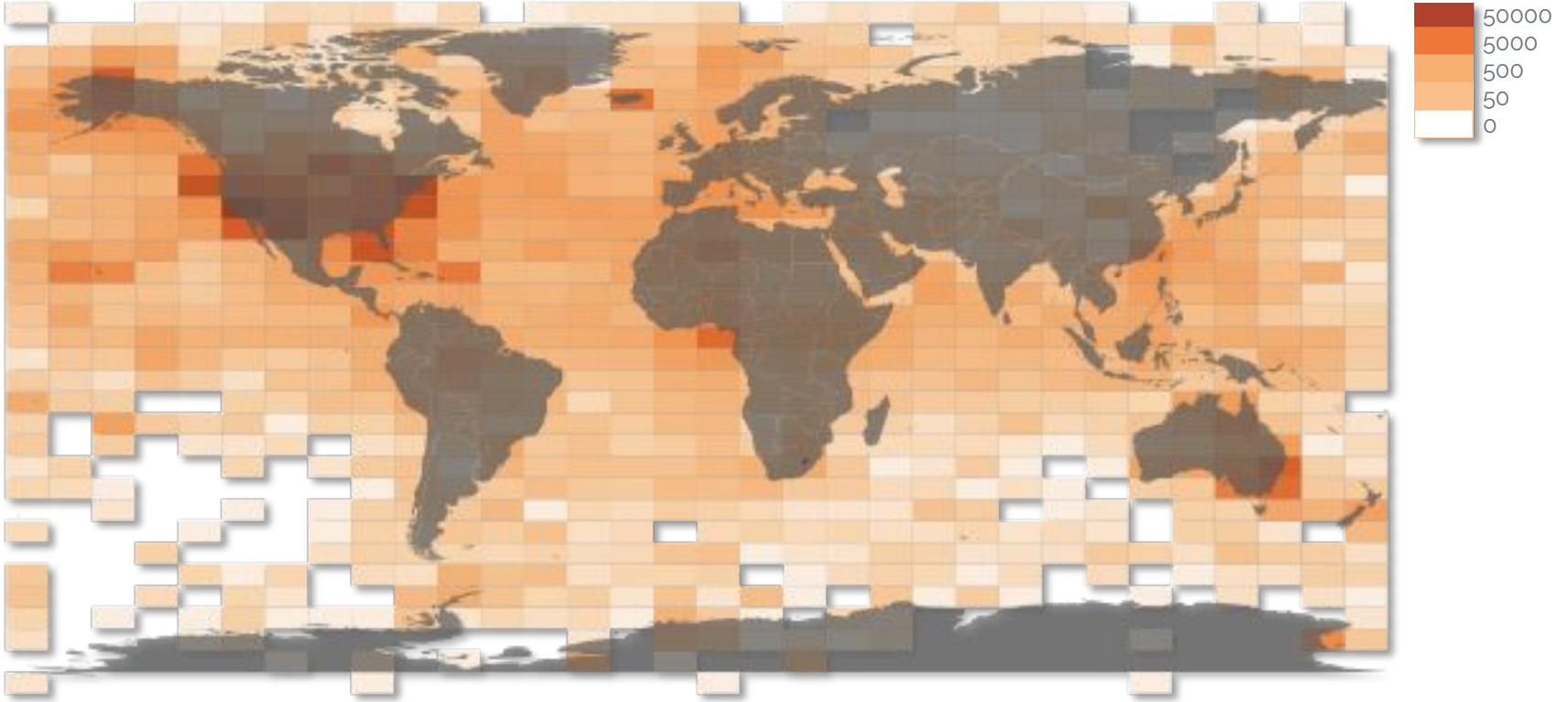
0000-0003-0077-4738
@metamattj

0000-0002-2192-403X
@sblifer

May 14, 2019

DataONE


Year
2012
2013
2014
2015
2016
2017
2018
Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenS



Global Data Coverage

Data Heterogeneity

Many Disciplines

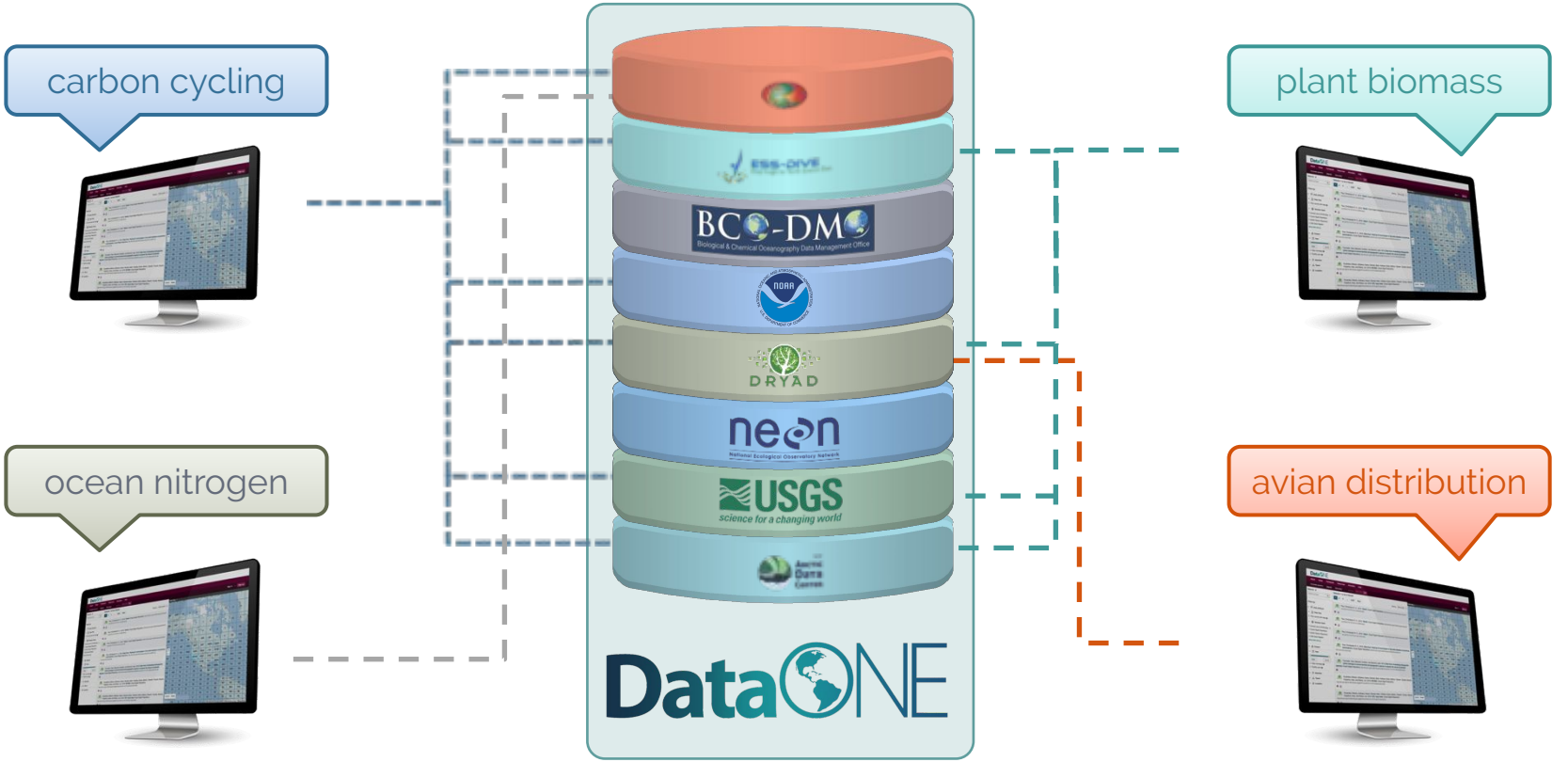
Communities Of Practice

Built Infrastructure

Drivers for loose coupling

Federated Search

Data Discovery and Access from Multiple Repositories



MetaDIG: Metadata Improvement and Guidance

NSF Grant to Habermann and Jones

Help scientific communities:

- Improve data discovery, and access
- Enable re-use
- Enhance understanding, especially across domains

… by improving metadata completeness and consistency through:

- Metadata evaluation and rubric design
- Metadata quality evaluation tools and services

The HDF Group

NCEAS

Target Audiences

Data producers: Individual researchers

- At record level, during submission

Data repositories

- At collection level

Consumers:

- *Individual researchers*
- *Repository managers*
- *Librarians*
- *User groups*
- *Funders*

Quality Improvement at Dataset Level

knb ABOUT DATA SUBMIT TOOLS Jump to: DOI or ID Go

‹ Back to search | Home / Search / Metadata

Melanie Frazier, Jamie Afflerbach, Casey O'Hara, Julia Steward Lowndes, Courtney Scarborough, et al. 2017. Ocean Health Index: 2016 Global. Knowledge Network for Biocomplexity. doi:10.5063/F1FX77DQ.

knb

Citations 0 | Downloads 0 | Views 0 | Copy Citation | Quality report

Files in this dataset Package: urn:uuid:e29f6af4-2882-4675-ae3f-2ba2904b77d6

Name		File type	Size	Download All
Metadata: Ocean Health Index: 2016 Global		EML v2.1.1	7 KB	Download
ohi-global_2016.zip	More info	octet stream (application file)	41 MB	Download

MetaDIG: Metadata Improvement and Guidance

Metadata Quality Report

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.

Quality suite: DataONE Metadata Completeness Suite v1.0

Identification: 88% complete

Discovery: 100% complete

Interpretation: 100% complete

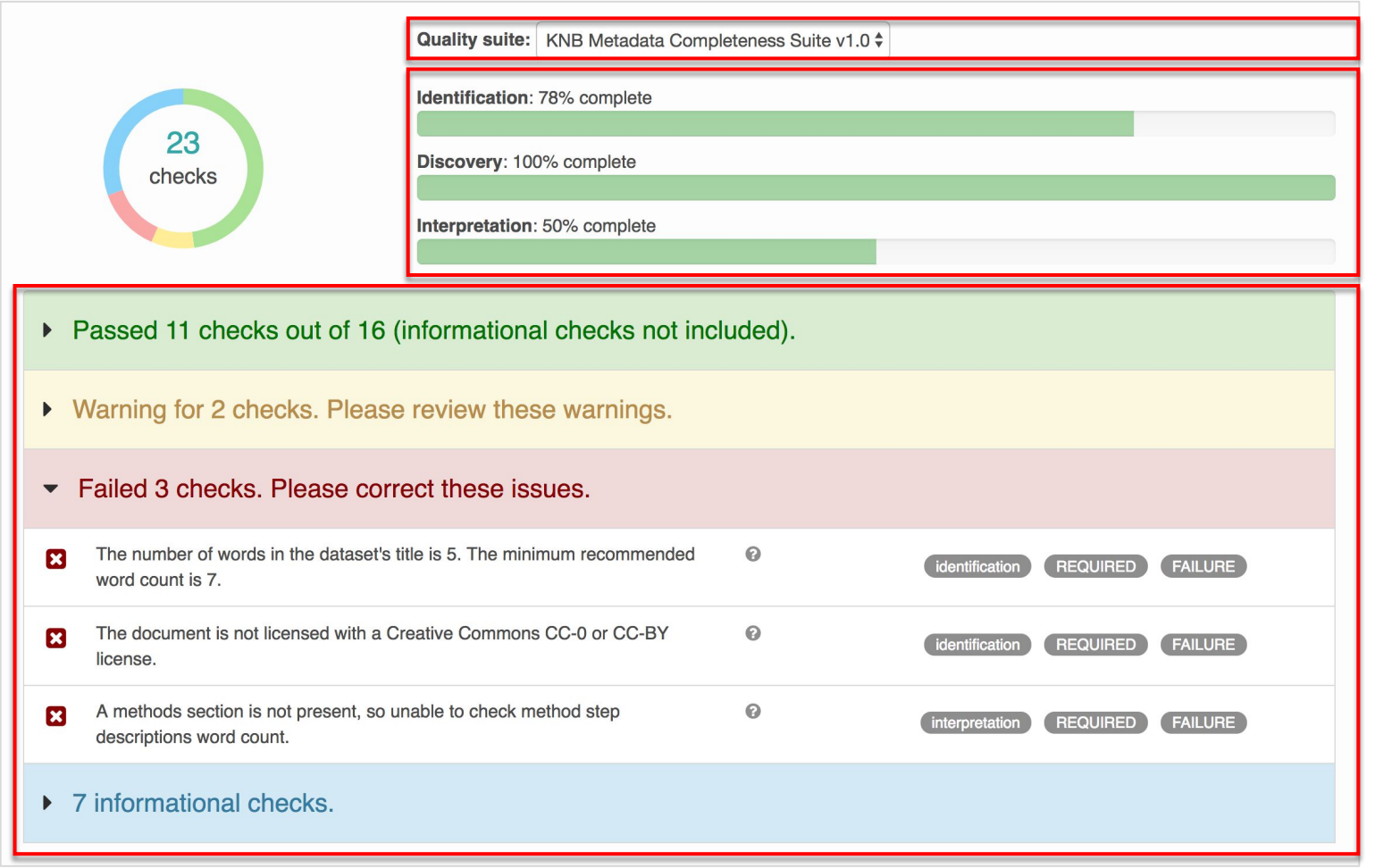
- **Passed 14 checks out of 20 (informational checks not included).**
- **Warning for 5 checks. Please review these warnings.**
- **Failed 1 check. Please correct these issues.**

More than one license was found which was an unexpected state. identification REQUIRED FAILURE

NCEAS

The HDF Group

Quality Improvement (cont.)



Extensible Quality Checks

Check#	Check Name	Check	Type
M1	Descriptive Title	Title exists, > 7 words	discovery
M2	Unique Attribute Names	Attribute names unique	discovery
M3	Valid Units	Units assigned from controlled vocabulary	interpretation
M4	Schema valid	Metadata validates	interpretation
C1	Checksum matches	Data checksums match metadata	reuse
C2	Data links live	All URLs return data	reuse
D1	Duplicate data rows	Count duplicate rows	reuse
…			

Recommendation -> Quality Suite

Suite is an implementation of a recommendation

- Contains a group of quality checks
- Can be created by any community
- Can include standard or custom checks
- Checks can access both metadata and data

Recommendation	Checks
LTER Best Practice	M1, M2, C2, C3, D3, …
ACDD	M2, M3, M4, C1, C2, D3, …
Arctic Data Center	M3, M4, M5, C6, C8, D1, D2, D3, …
…	

Current MetaDIG suites

- Arctic Data Center
- KNB Data Repository Suite
- ESS-DIVE Repository

- **DataONE FAIR Suite**

Quantifying FAIR, 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

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 they can they be accessed, possibly including authentication and authorization.

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.

Findable

Findable Guidelines

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 they describe

F4. (Meta)data are registered or indexed in a searchable resource

Accessible

Accessible Guidelines

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

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

Interoperable

Interoperable Guidelines

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

Reusable

Reusable Guidelines

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

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

FAIR Metrics workshop March 2019

- Matt Jones, Ted Habermann, Sean Gordon, Peter Slaughter, Amber Budden, Daniella Lowenberg, John Chodacki, Margaret O'Brien

FAIR Metrics

Findable

DataONE FAIR Checks

Item that is checked	Description of check	Facet	Required	Implemented
title	presence, length, content	F2	Y	partially
metadata identifier	presence, identifier type	F1	Y	partially
resource identifier	presence, identifier 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

Findable

DataONE FAIR Checks

Item that is checked	Description of check	Facet	Required	Implemented
spatial coverage	presence	F2	N	Y
taxonomic coverage	presence	F2	N	Y
natural language keywords	presence	F2	N	Y
controlled keywords	presence	F2	N	Y
creator/author	presence, email, identification, affiliation	F2	Y	partially
creator/author identifier	presence	F2	Y	Y

Accessible

DataONE FAIR 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

Accessible

DataONE FAIR Checks

Item that is checked	Description of check	Facet	Required	Implemented
data are public	public user allowed	A1.2	N	N
authenticated access	users, user groups allowed	A1.2	N	N

Interoperable

DataONE FAIR 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 definitions	presence	I2	Y	Y
attribute names unique	for an entity, names are unique	I2	Y	N
attribute storage type	presence	I2	Y	Y

Interoperable

DataONE FAIR Checks

Item that is checked	Description of check	Facet	Required	Implemented
landing page uses schema.org	inspect landing page	I2	Y	N

Reusable

DataONE FAIR 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, congruence	R1.3	Y	partially

Reusable

DataONE FAIR Checks

Item that is checked	Description of check	Facet	Required	Implemented
attribute precision	presence	R1.3	Y	N
provenance process steps	presence	R1.2	Y	Y
provenance sources	presence	R1.2	Y	Y
provenance is PROV-O	check content	R1.2	Y	N
quality data	presence	R1.3	Y	N
citation for reuse attribution	presence		Y	N

DataONE FAIR checks

Source code for checks is available at https://github.com/NCEAS/metadig-checks

Community contributions are welcome!

DataONE Infrastructure Services

DataONE Quality Assessment Service

A scalable approach to metadata evaluation

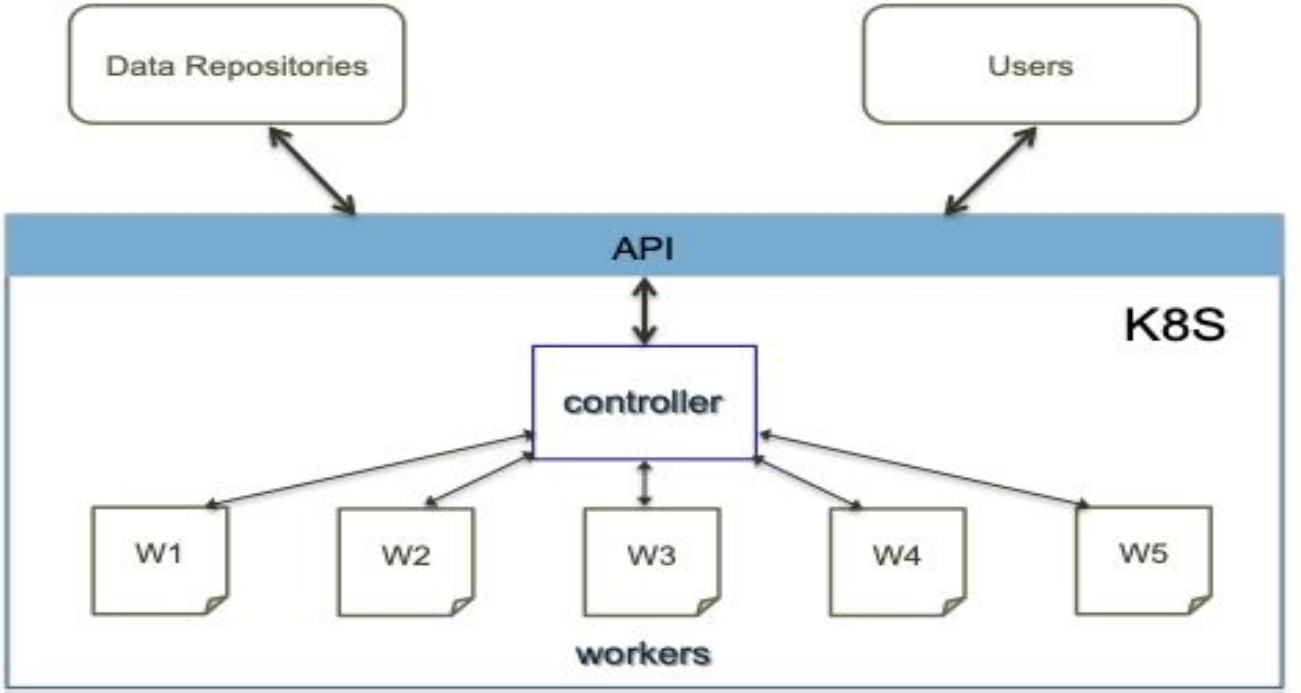
Problem: Need to process millions of metadata record versions, for many different quality suites, repeatedly as suites change

- Automatic queueing system detects new metadata versions
- The DataONE quality service can be scaled to hundreds or thousands of cores
  - Enabled via a Kubernetes compute cluster

kubernetes

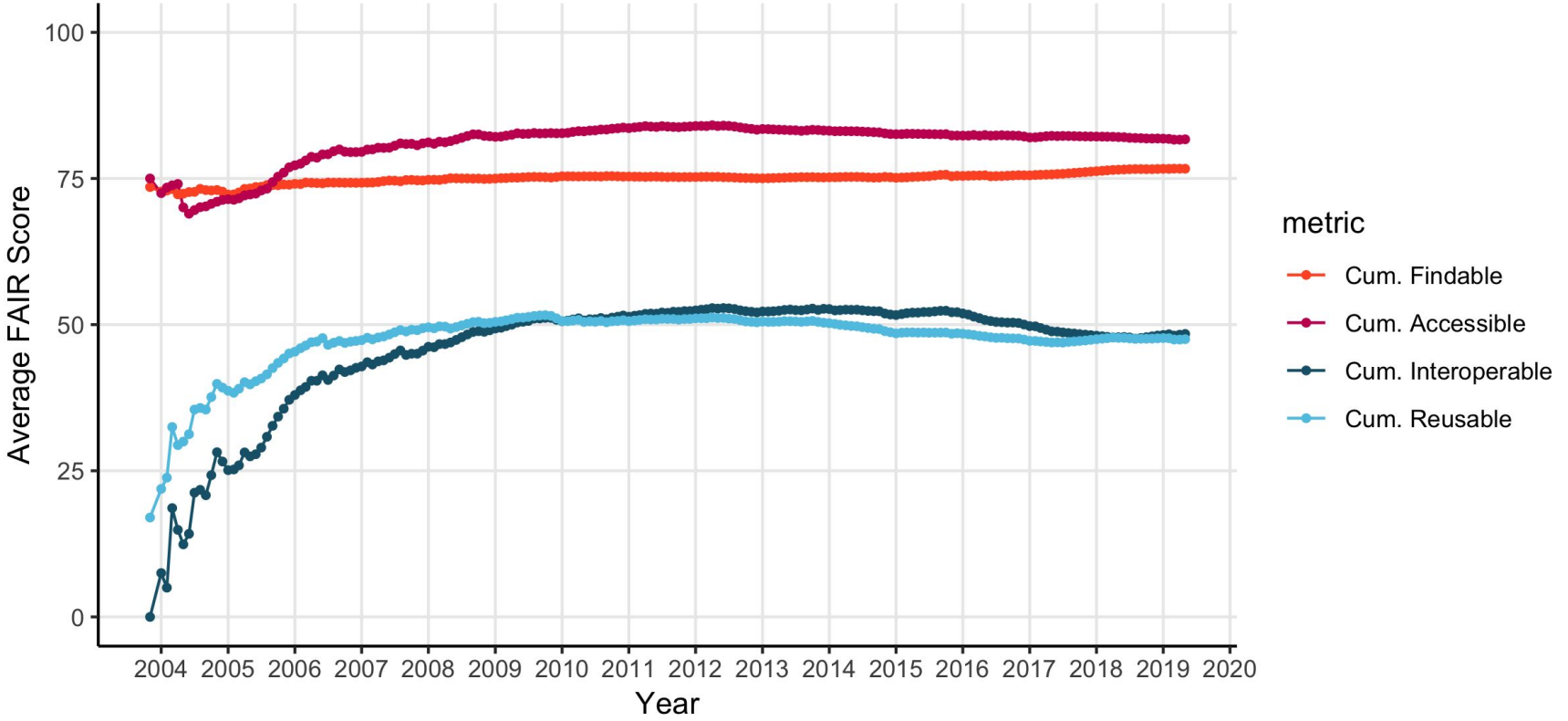
DataONE Quality Assessment Service

A scalable approach to metadata evaluation (cont)

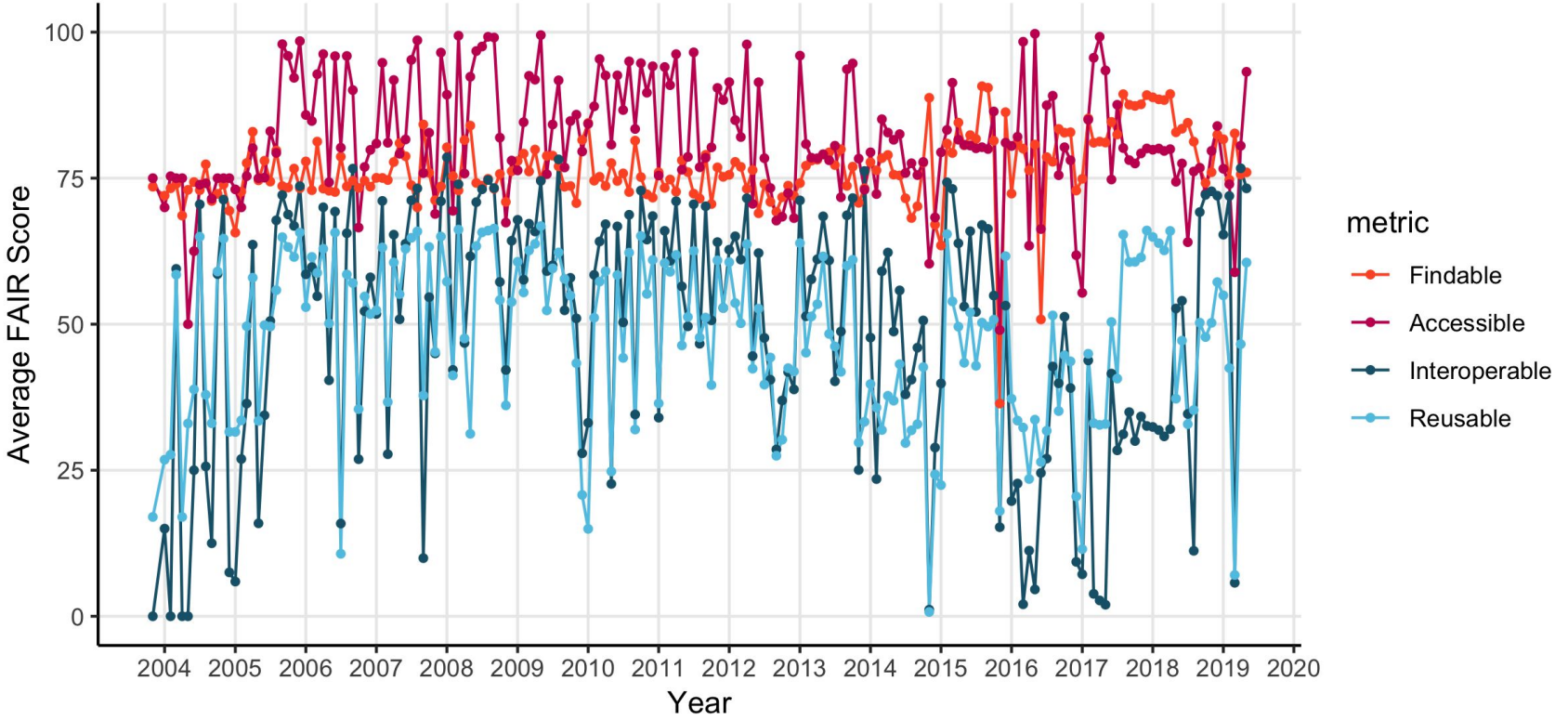


Are datasets in DataONE FAIR? Preliminary results

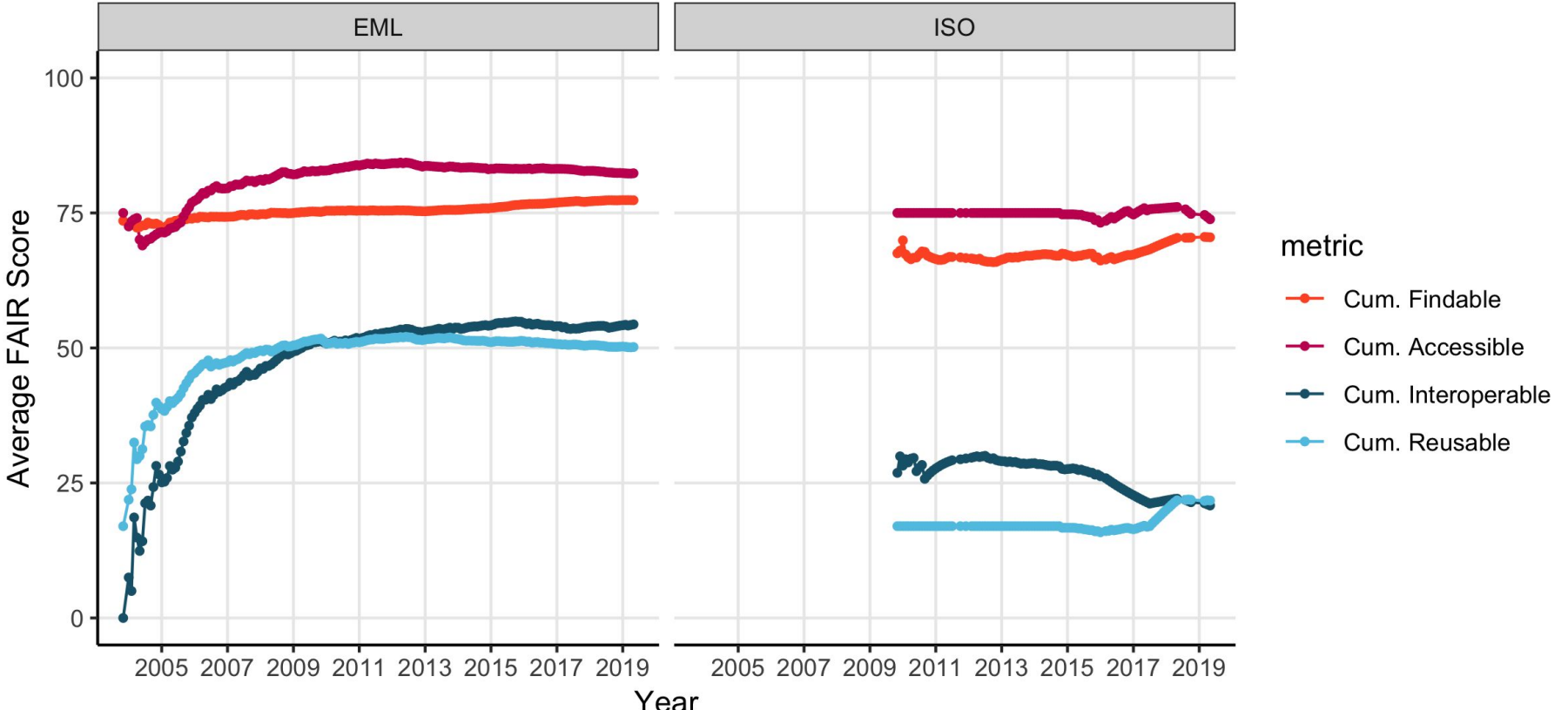
DataONE: FAIR scores for 687,126 EML and ISO metadata records



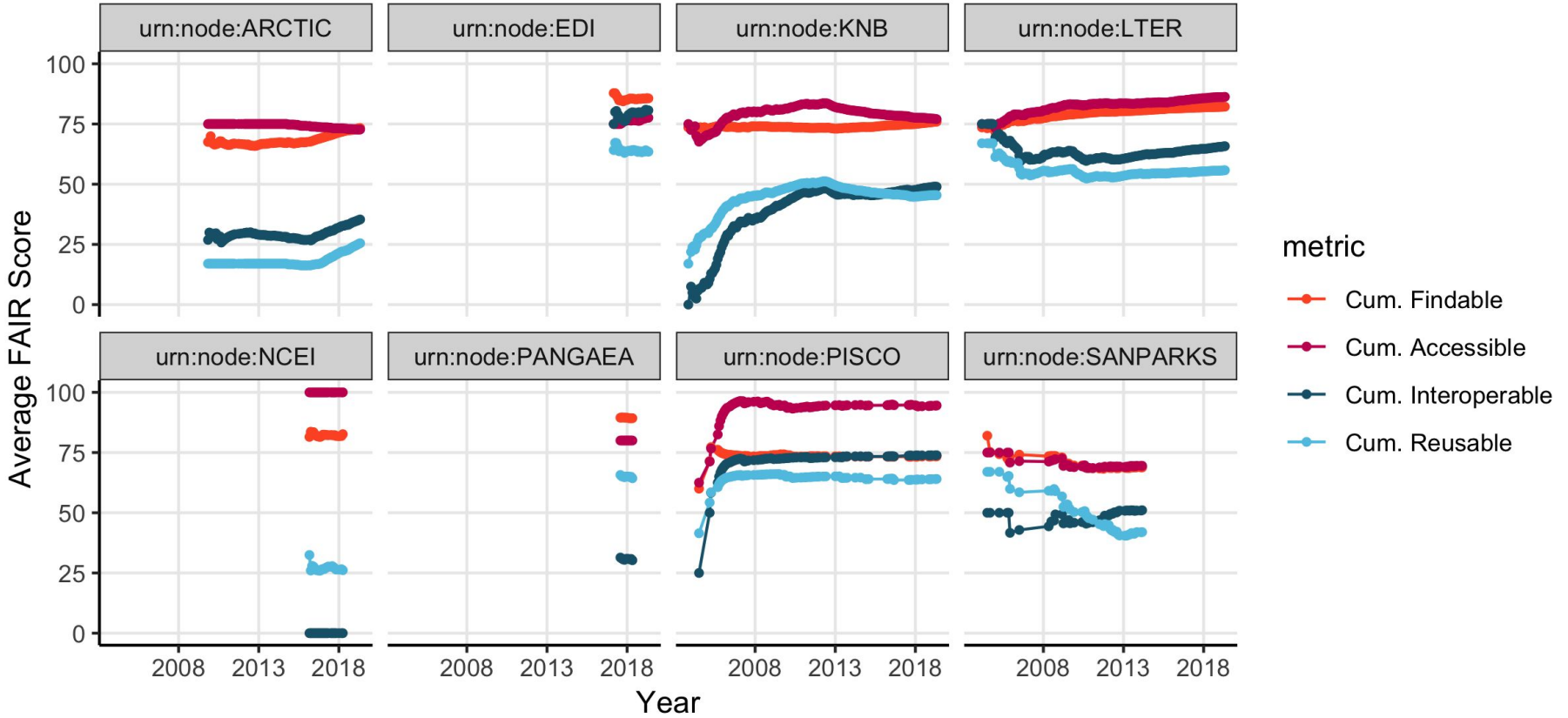
DataONE: FAIR scores for 687,126 EML and ISO metadata records



DataONE: FAIR scores for 119,913 EML and 567,213 ISO metadata records



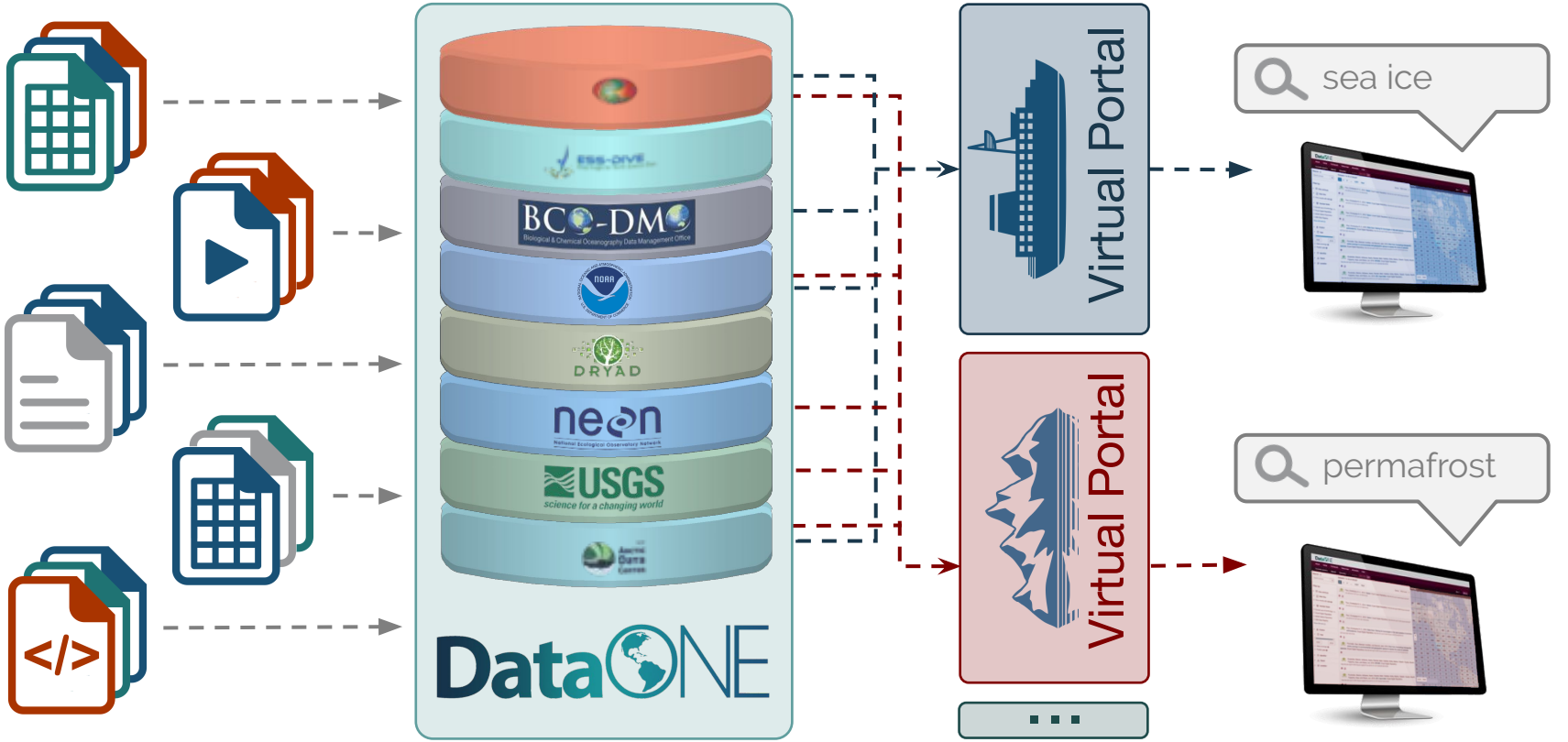
DataONE: FAIR scores for selected repositories



DataONE Metadata Quality Services

Virtual Portals

Data Aggregated from Repositories Across DataONE



Virtual Portal Services

Customizable Services by Organization, Theme and Region

Theme

Project/ Organization

Region

Repository

Custom branding and storytelling

Data Level Metrics

Aggregated views, downloads and citations

Metadata Quality

Customized checks of metadata quality

Metadata Creation

Metadata authoring and data upload tools

Customized Search

Community specific search features for enhanced discovery

...

Virtual Portal Services

Customizable Services by Organization, Theme and Region

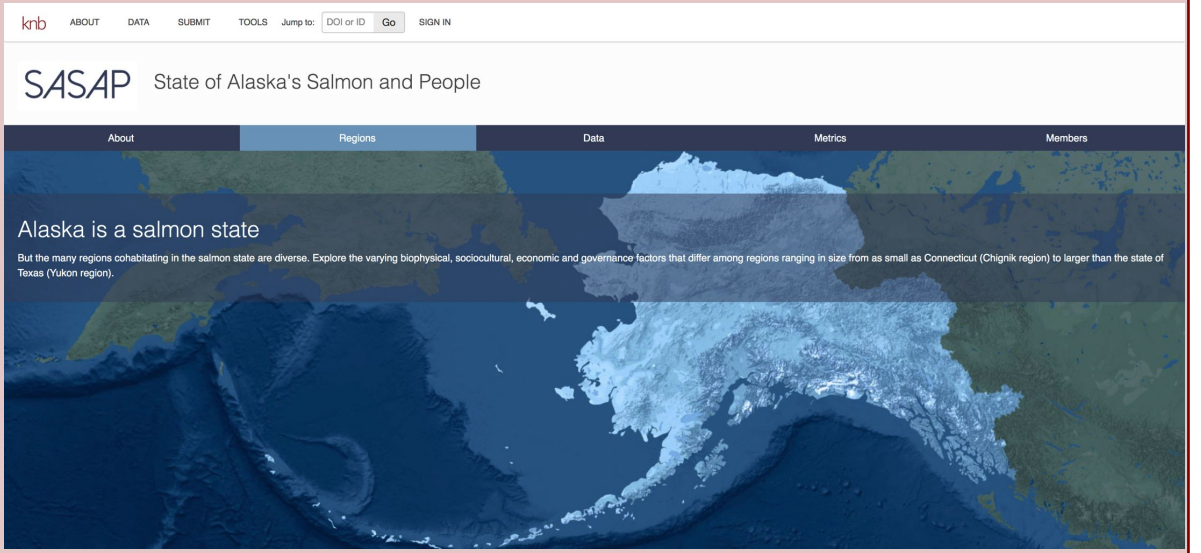
Theme

Project/ Organization

Region

Repository

Custom branding and storytelling



Virtual Portal Services

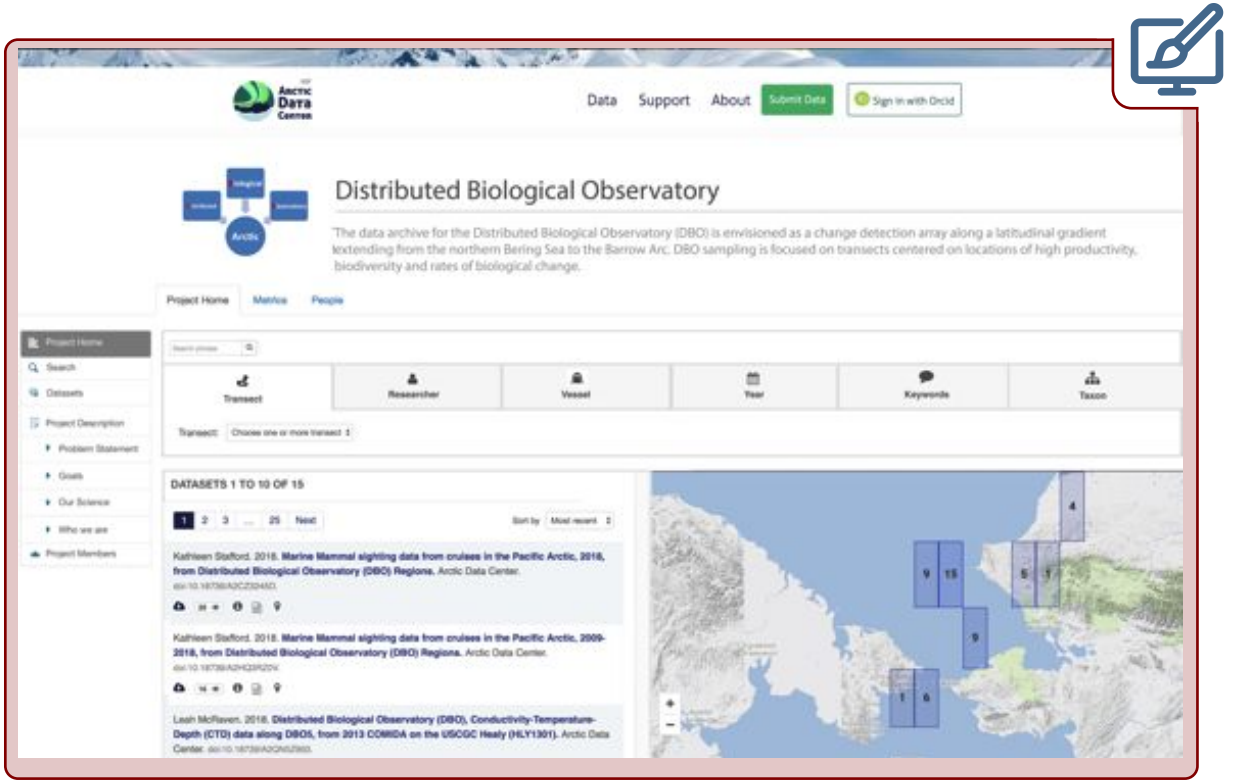
Customizable Services by Organization, Theme and Region

Theme

Project/ Organization

Region

Repository

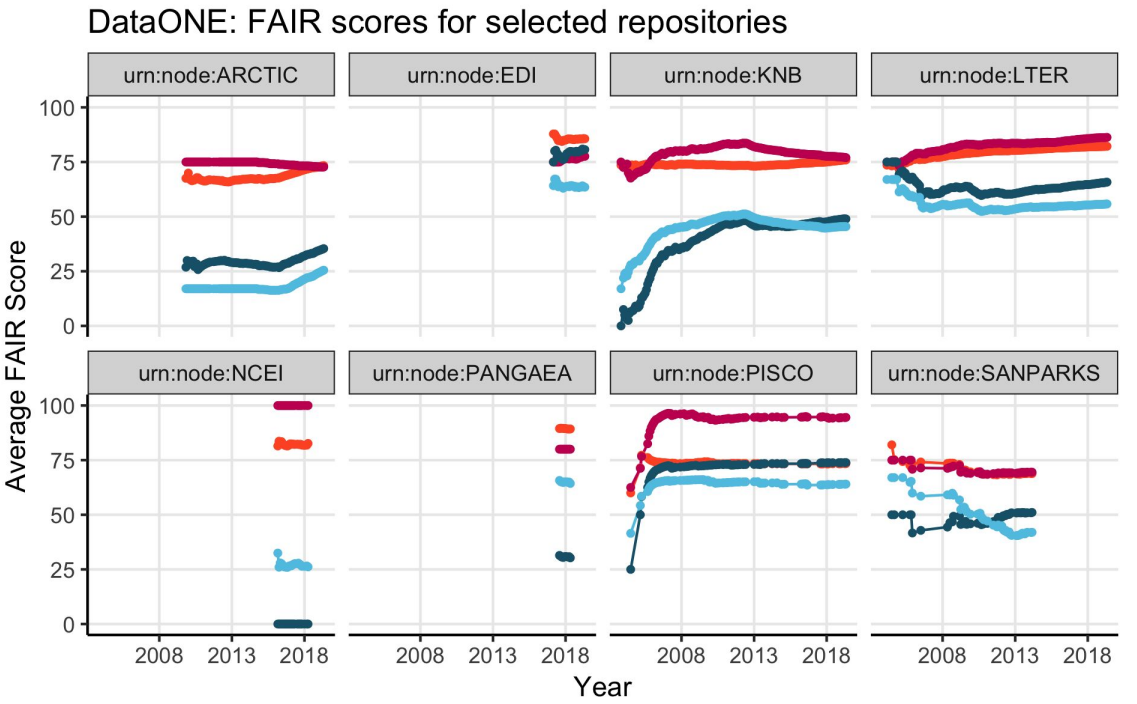


Longitudinal Quality Reports

FAIR and custom quality reports aggregated by:

- Single repository
- Entire network
- User
- User group
- Project

Custom guidance and consulting services



Big thanks to our collaborators:

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

DataONE