Sharing Data Through Guided Metadata Improvement

Lindsay Powers and **Ted Habermann** - The HDF Group

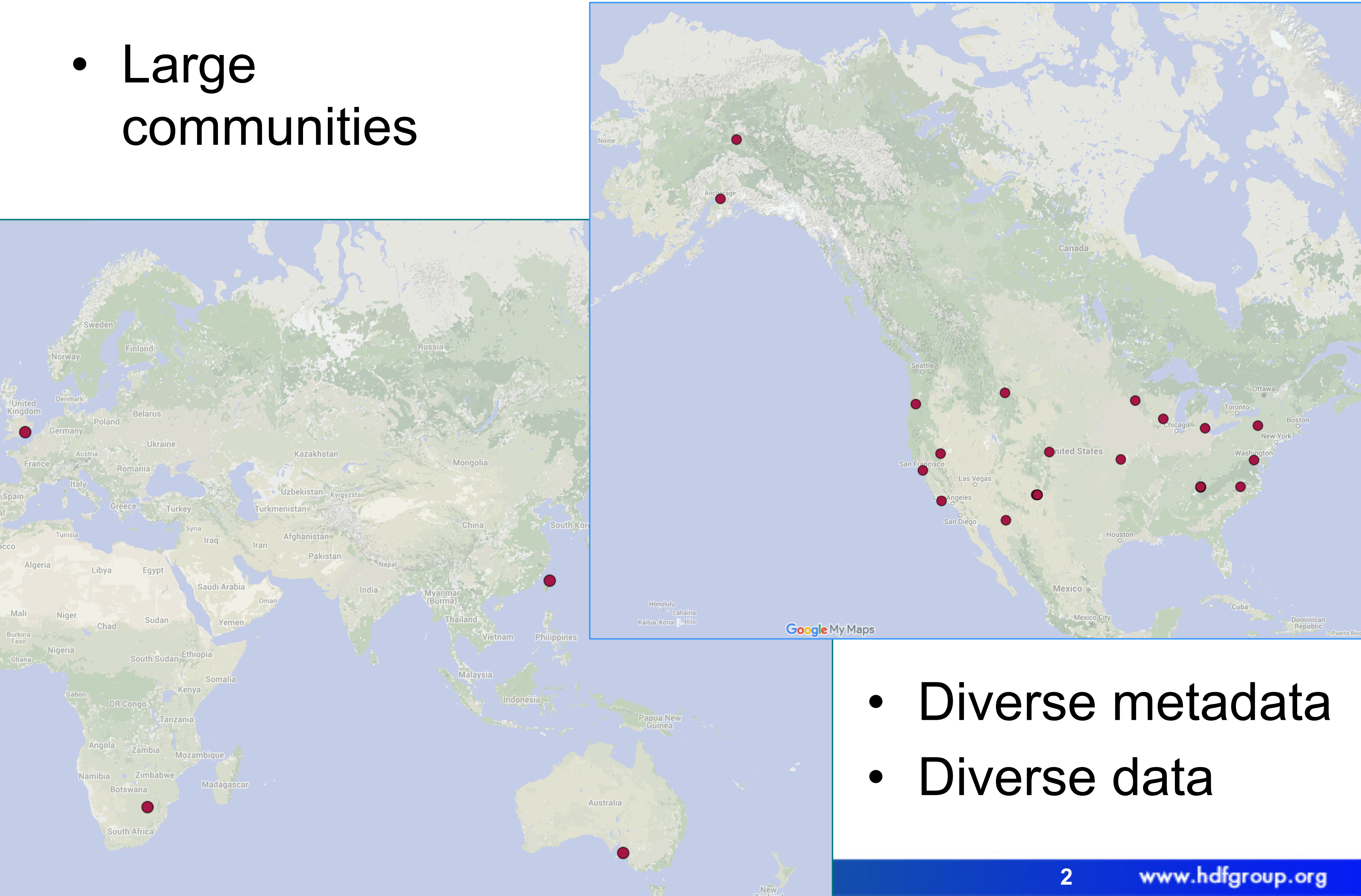
Matthew B. Jones – National Center for Ecological Analysis and Synthesis, University of California Santa Barbara

NCEAS
National Center for Ecological Analysis and Synthesis

DataONE
Data Observation Network for Earth

DataONE Repositories

- Large communities



- Diverse metadata
- Diverse data

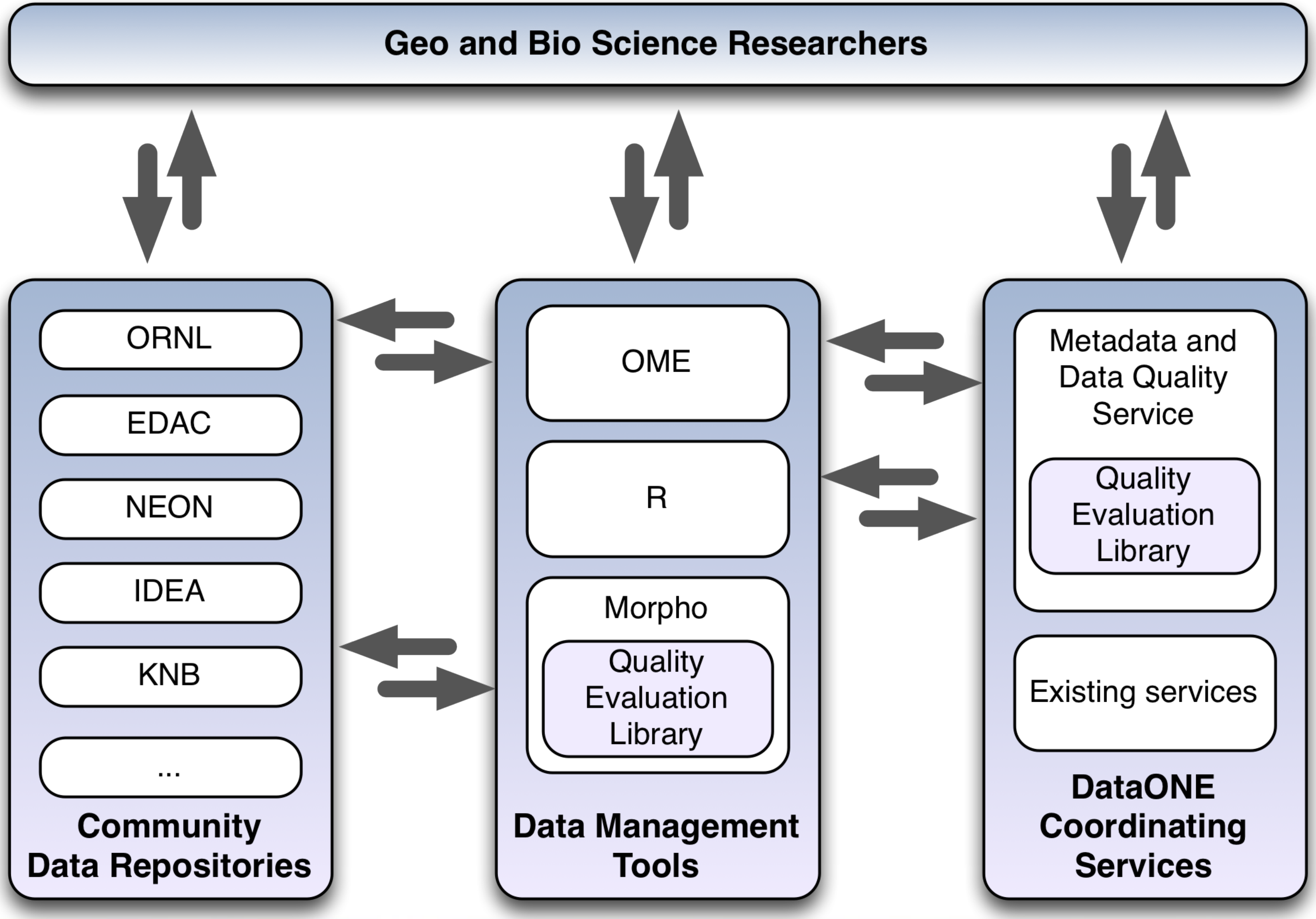
MetaDIG Tools and services

- Metadata Improvement and Guidance (MetaDIG):
  - Individual researchers (producers)
    - At record level, during submission
  - Data repositories
    - At collection level
  - Individual researchers (consumers)
    - At record level, for re-use

Automation

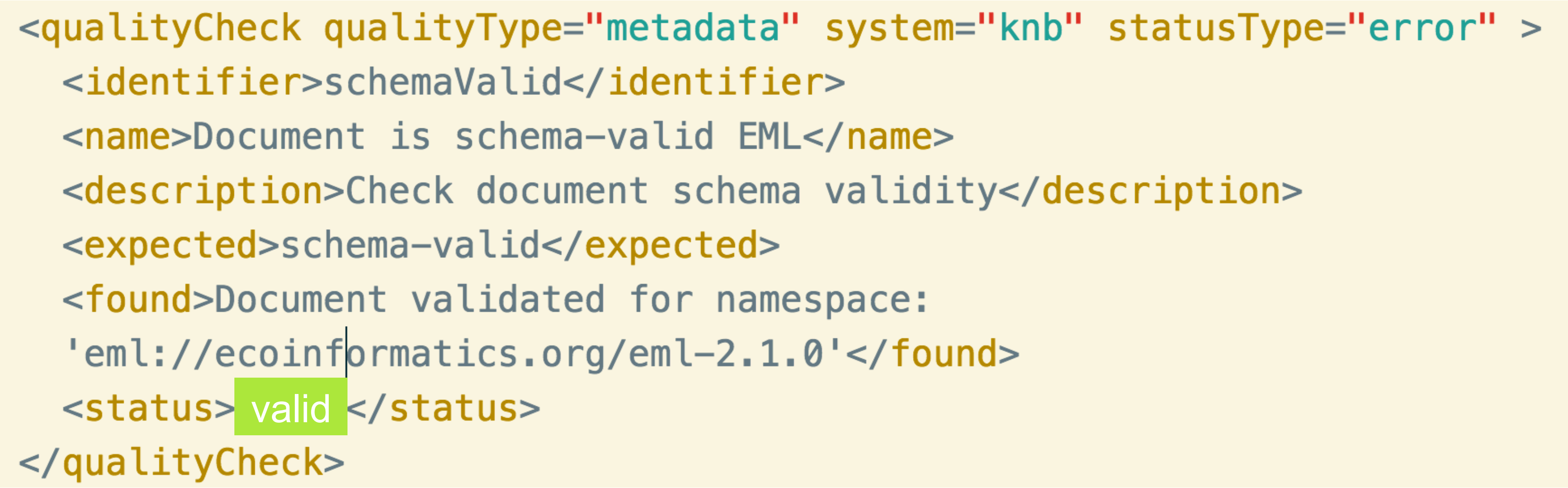
- Automate:
  - Metadata **Completeness**
    - against recommendations
  - Metadata and Data **Congruency**
  - Metadata **Effectiveness**
    - Semantics, therefore much harder

Metadata Quality Service


Geo and Bio Science Researchers
ORNL
EDAC
NEON
IDEA
KNB
...
Community Data Repositories
OME
R
Morpho
Quality Evaluation Library
Data Management Tools
Metadata and Data Quality Service
Quality Evaluation Library
Existing services
DataONE Coordinating Services

EML Congruency Checker

- Starting point:
  - LTER tool for Ecological Metadata Language
  - Standard, extensible report format
  - Suite of developed checks



```
<qualityCheck qualityType="metadata" system="knb" statusType="error" >
  <identifier>schemaValid</identifier>
  <name>Document is schema-valid EML</name>
  <description>Check document schema validity</description>
  <expected>schema-valid</expected>
  <found>Document validated for namespace:
  'eml://ecoinformatics.org/eml-2.1.0'</found>
  <status> valid </status>
</qualityCheck>
```

Extensible quality checks

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

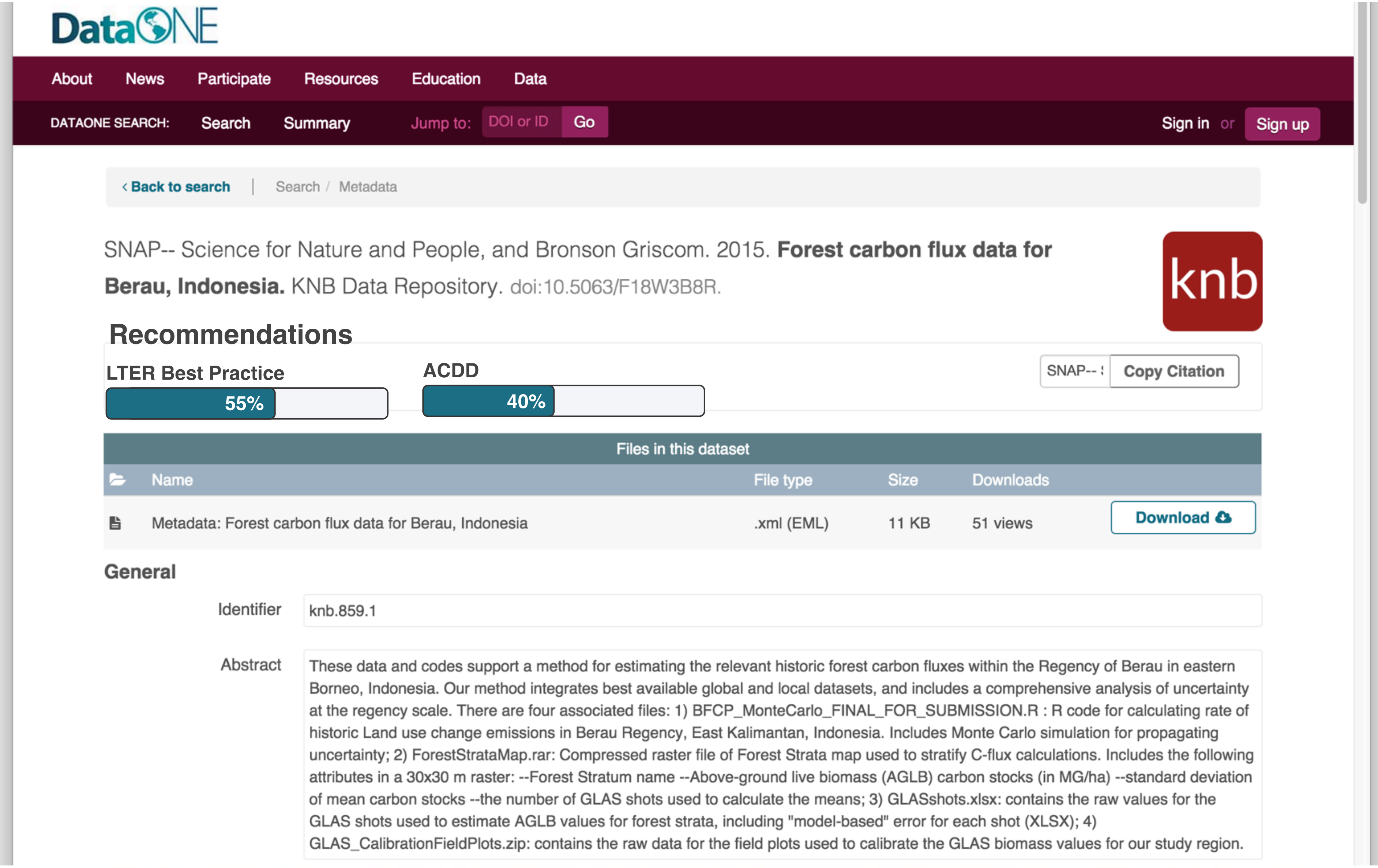
- Checks in Java, R, Python
- Categorized by function (discovery, re-use, …)
- Operate across dialects (EML, CSDGM, ISO19139)

Recommendations

- Checks: like unit tests for recommendations
- Community Recommendations
  - Group of quality checks
  - Can be created by any community
  - Can include standard or custom checks
  - Checks: access both metadata and data

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

For Creators



SNAP-- Science for Nature and People, and Bronson Griscom. 2015. **Forest carbon flux data for Berau, Indonesia.** KNB Data Repository. doi:10.5063/F18W3B8R.

Recommendations

LTER Best Practice 55%

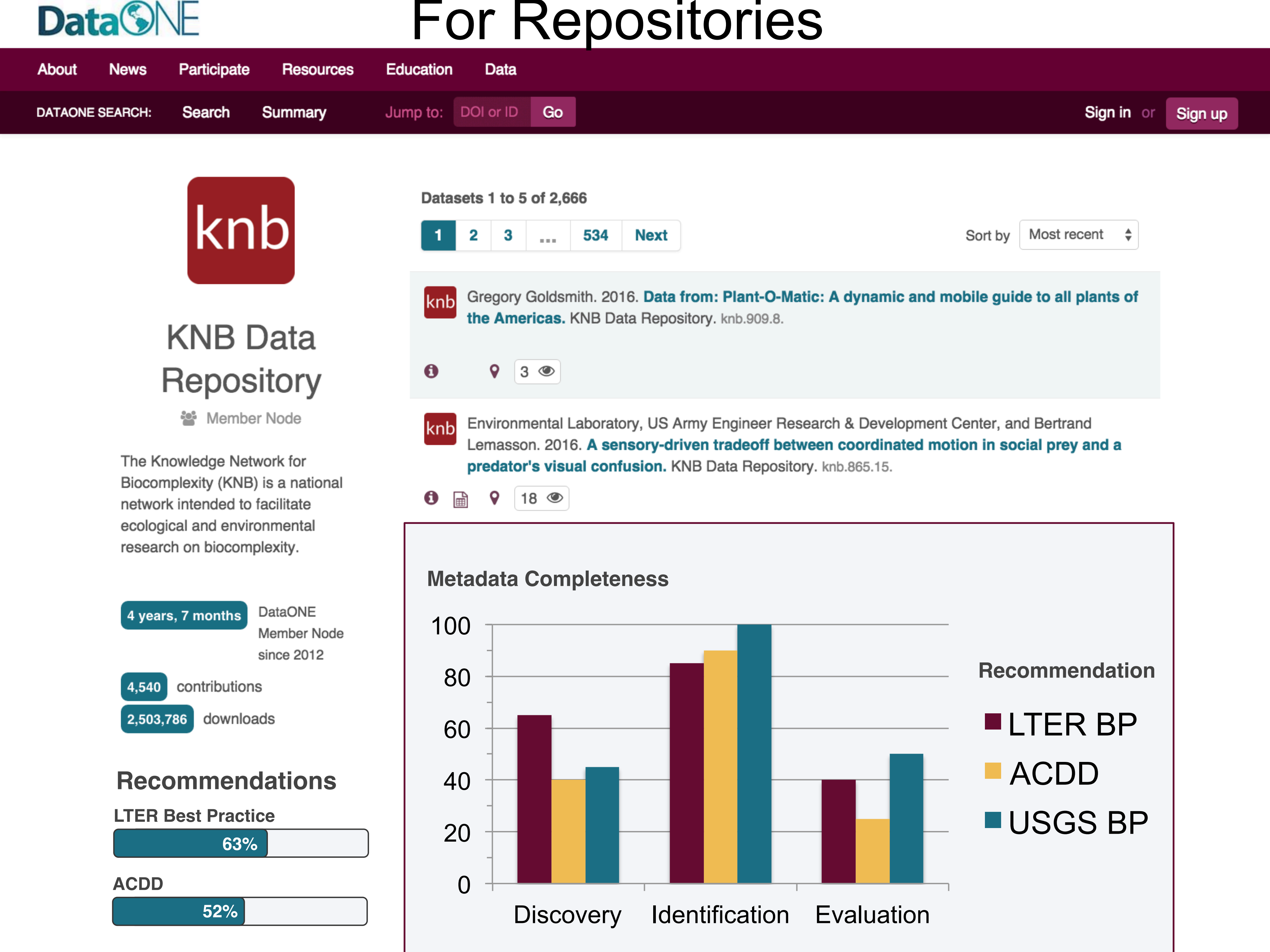
ACDD 40%

Files in this dataset				
Name	File type	Size	Downloads	
Metadata: Forest carbon flux data for Berau, Indonesia	.xml (EML)	11 KB	51 views	Download

General

Identifier	knb.859.1
Abstract	These data and codes support a method for estimating the relevant historic forest carbon fluxes within the Regency of Berau in eastern Borneo, Indonesia. Our method integrates best available global and local datasets, and includes a comprehensive analysis of uncertainty at the regency scale. There are four associated files: 1) BFCP_MonteCarlo_FINAL_FOR_SUBMISSION.R : R code for calculating rate of historic Land use change emissions in Berau Regency, East Kalimantan, Indonesia. Includes Monte Carlo simulation for propagating uncertainty; 2) ForestStrataMap.rar: Compressed raster file of Forest Strata map used to stratify C-flux calculations. Includes the following attributes in a 30x30 m raster: --Forest Stratum name --Above-ground live biomass (AGLB) carbon stocks (in MG/ha) --standard deviation of mean carbon stocks --the number of GLAS shots used to calculate the means; 3) GLASshots.xlsx: contains the raw values for the GLAS shots used to estimate AGLB values for forest strata, including "model-based" error for each shot (XLSX); 4) GLAS_CalibrationFieldPlots.zip: contains the raw data for the field plots used to calibrate the GLAS biomass values for our study region.

For Repositories



DataONE

About News Participate Resources Education Data

DATAONE SEARCH: Search Summary Jump to: DOI or ID Go Sign in or Sign up

knb

KNB Data Repository

Member Node

The Knowledge Network for Biocomplexity (KNB) is a national network intended to facilitate ecological and environmental research on biocomplexity.

4 years, 7 months DataONE Member Node since 2012

4,540 contributions

2,503,786 downloads

Recommendations

LTER Best Practice

63%

ACDD

52%

Datasets 1 to 5 of 2,666

1 2 3 ... 534 Next Sort by Most recent

Gregory Goldsmith. 2016. **Data from: Plant-O-Matic: A dynamic and mobile guide to all plants of the Americas.** KNB Data Repository. knb.909.8.

3

Environmental Laboratory, US Army Engineer Research & Development Center, and Bertrand Lemasson. 2016. **A sensory-driven tradeoff between coordinated motion in social prey and a predator's visual confusion.** KNB Data Repository. knb.865.15.

18

Metadata Completeness

Recap

- MetaDIG project plans
  - Metadata evaluation and completeness
  - Metadata completeness tools and services
  - Communication, guidance, and outreach

Thanks

This work was supported by National Science Foundation award ACI - 1443062.