Sharing Data Through Guided Metadata Improvement

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

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

Goals

To help scientific communities:

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

… by improving metadata completeness and consistency.

Terminology

Collection: A group of metadata records ideally in a machine-readable format, commonly organized by a data center, organization or project and often stored in a database or web accessible folder.

Dialect : A particular form of the documentation language that is specific to a community (e.g. DIF, CSDGM, EML, ECHO, custom).

Concept : General term for describing a documentation entity (e.g. Title, Revision Date, Process Step, Spatial Extent).

Spiral: A set of concepts required to support a particular documentation need or use case.

Recommendation: A set of concepts that a group believes is required for achieving a documentation goal.

DataONE Member Nodes (communities) using EML*

- Ecological Society of America (**ESA**)
- Global Lake Ecological Observatory Network (**GLEON**)
- Alaska Ocean Observing System (**GOA**)
- Montana Institute on Ecosystems (**IOE**)
- Knowledge Network for Biocomplexity (**KNB**)
- University of Kansas Biodiversity Institute (**KUBI**)
- Long-term Ecological Research Network (**LTER**)
- European Long-term Ecosystem Research Network (**LTER_EUROPE**)
- University of California / DataONE (**ONEShare**)
- Terrestrial Environmental Research Network (**TERN**)
- Taiwan Forestry Research Institute (**TFRI**)
- National Phenology Network (**USANPN**)

** Ecological Metadata Language*

DataONE Member Nodes using CSDGM*

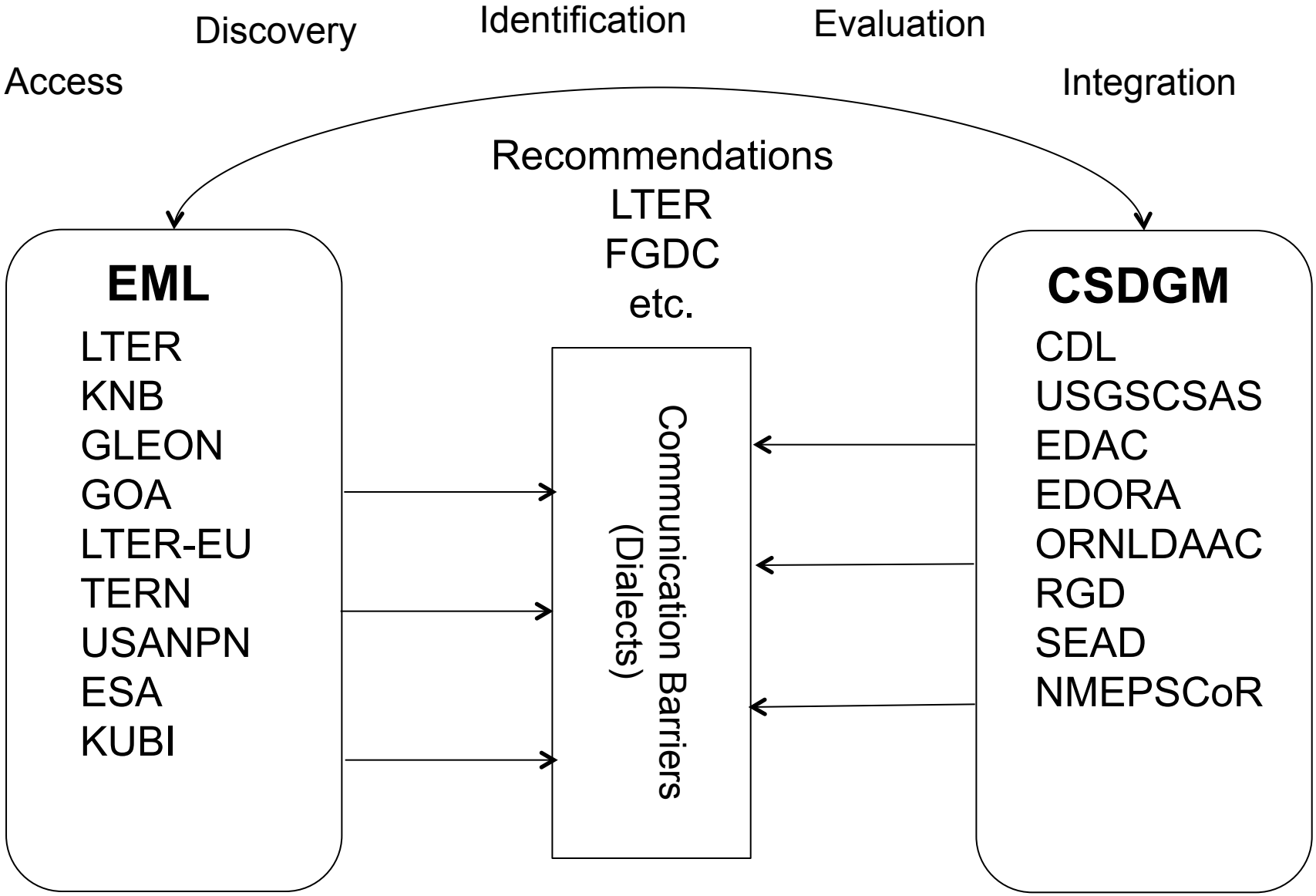
- California Digital Libraries (**CDL**)
- USGS Core Sciences Clearinghouse (**USGSCSAS**)
- Earth Data Analysis Center (**EDACGSTORE**)
- Environmental Data for the Oak Ridge Area (**EDORA)**
- Oak Ridge National Lab Distributed Active Archive Center (**ORNLDAAC**)
- Regional and Global Biogeochemical Dynamics Data (**RGD**)
- Sustainable Environment Actionable Data (**SEAD**)
- New Mexico Experimental Program to Stimulate Comptetitive Research (**NMEPSCOR**)

** Content Standard for Digital Geospatial Metadata*

Questions

- Can community developed metadata recommendations help improve metadata content within a particular community?
- Can community developed metadata recommendations help improve metadata content among different communities?
- Can metadata recommendations developed in a specific dialect be used to help improve metadata in other dialects? Can they facilitate communication?

Communities, dialects and recommendations

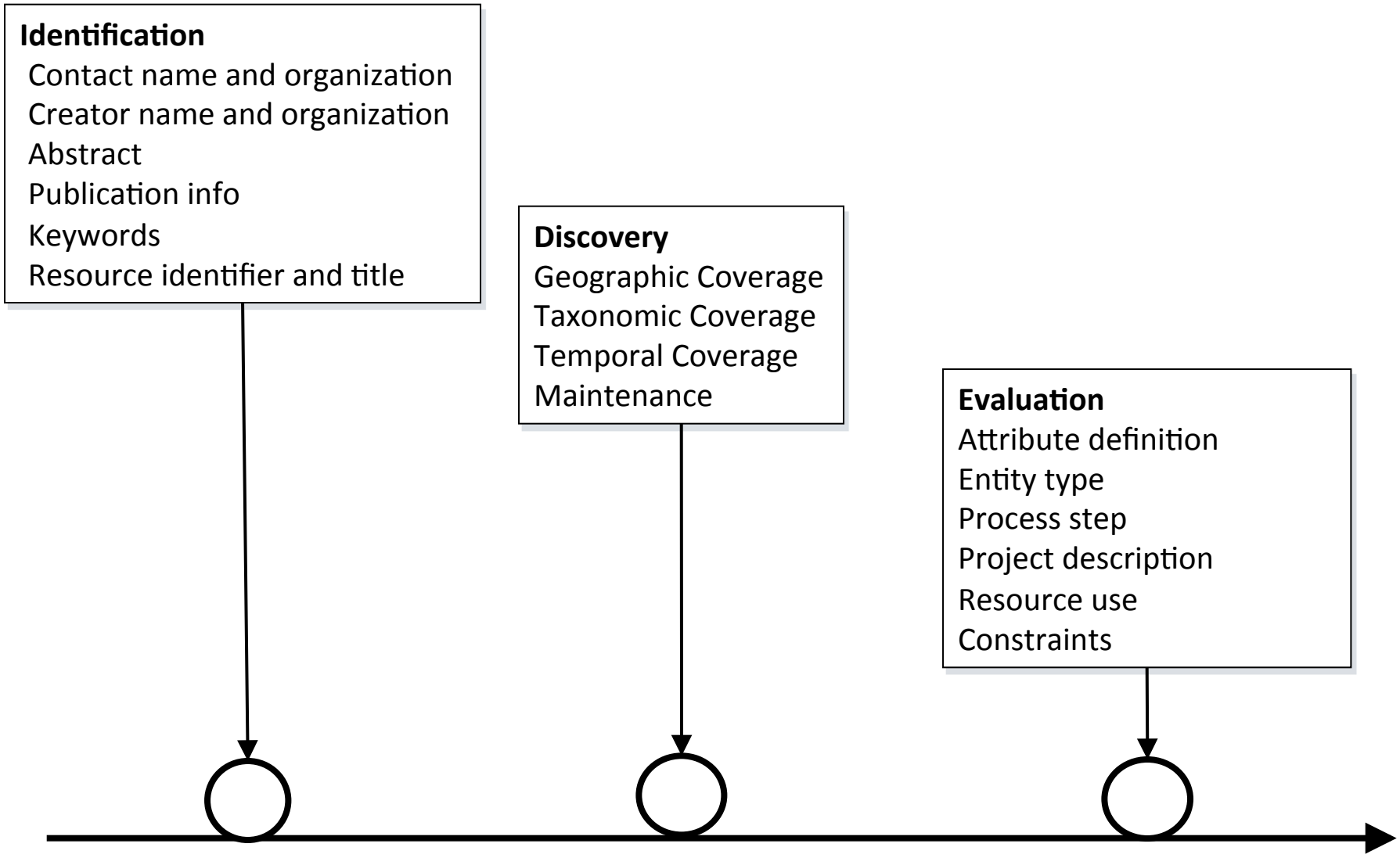


LTER Metadata Recommendations

LTER developed a suite of metadata recommendations based on community requirements…

- Did these recommendations make metadata more complete in comparison to other entities?
- Does LTER metadata practice provide a good example for other communities?

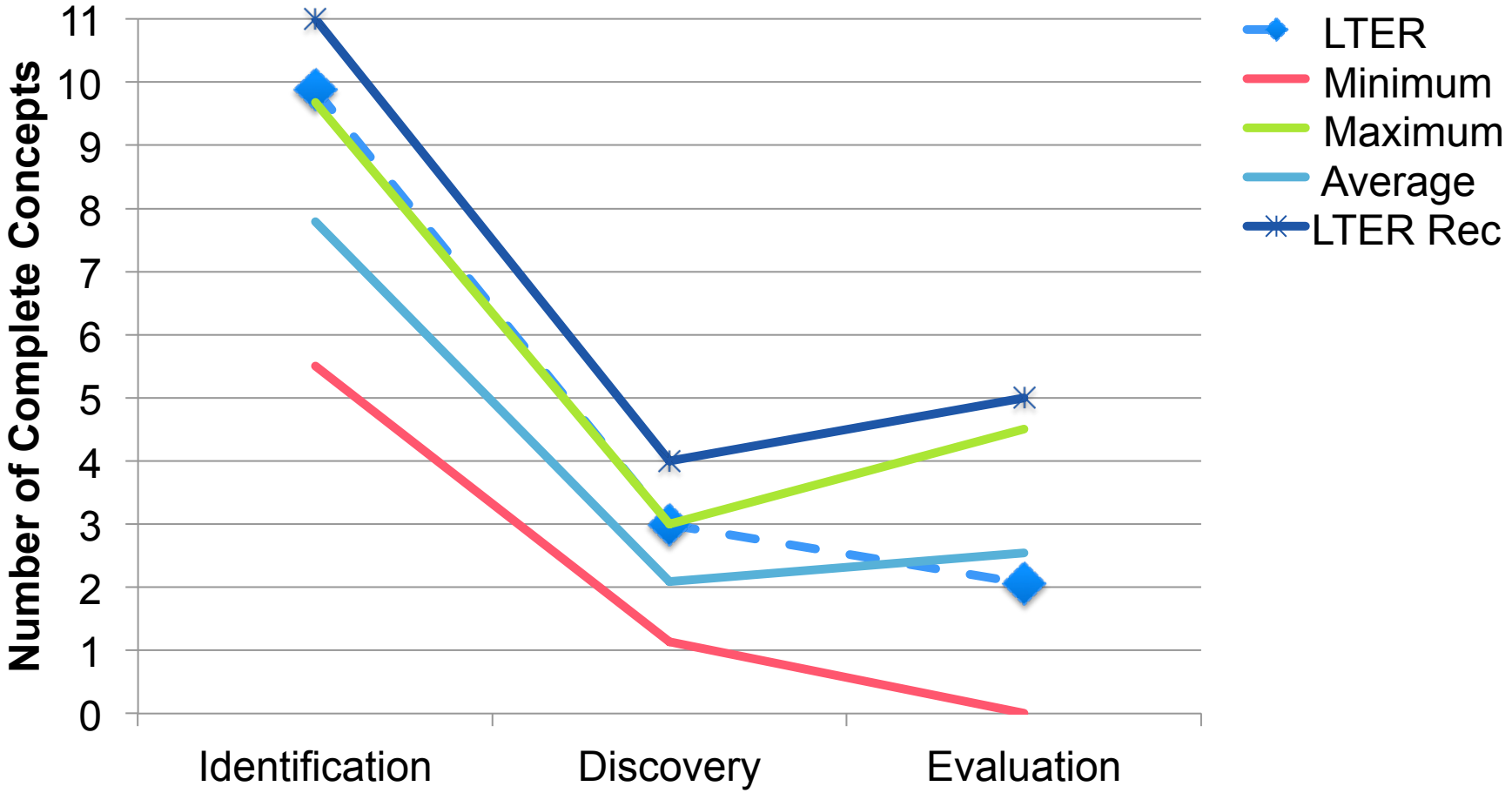
LTER Recommendations (Spirals)



Methods

- Randomly sampled up to 250 records from each EML or CSDGM metadata collection (Member Node)
- Mapped dialect to LTER Recommendation concepts
- Analyzed collections for completeness in relation to recommendations
- Compared collections to identify shining examples

LTER Recommendations and EML Collections

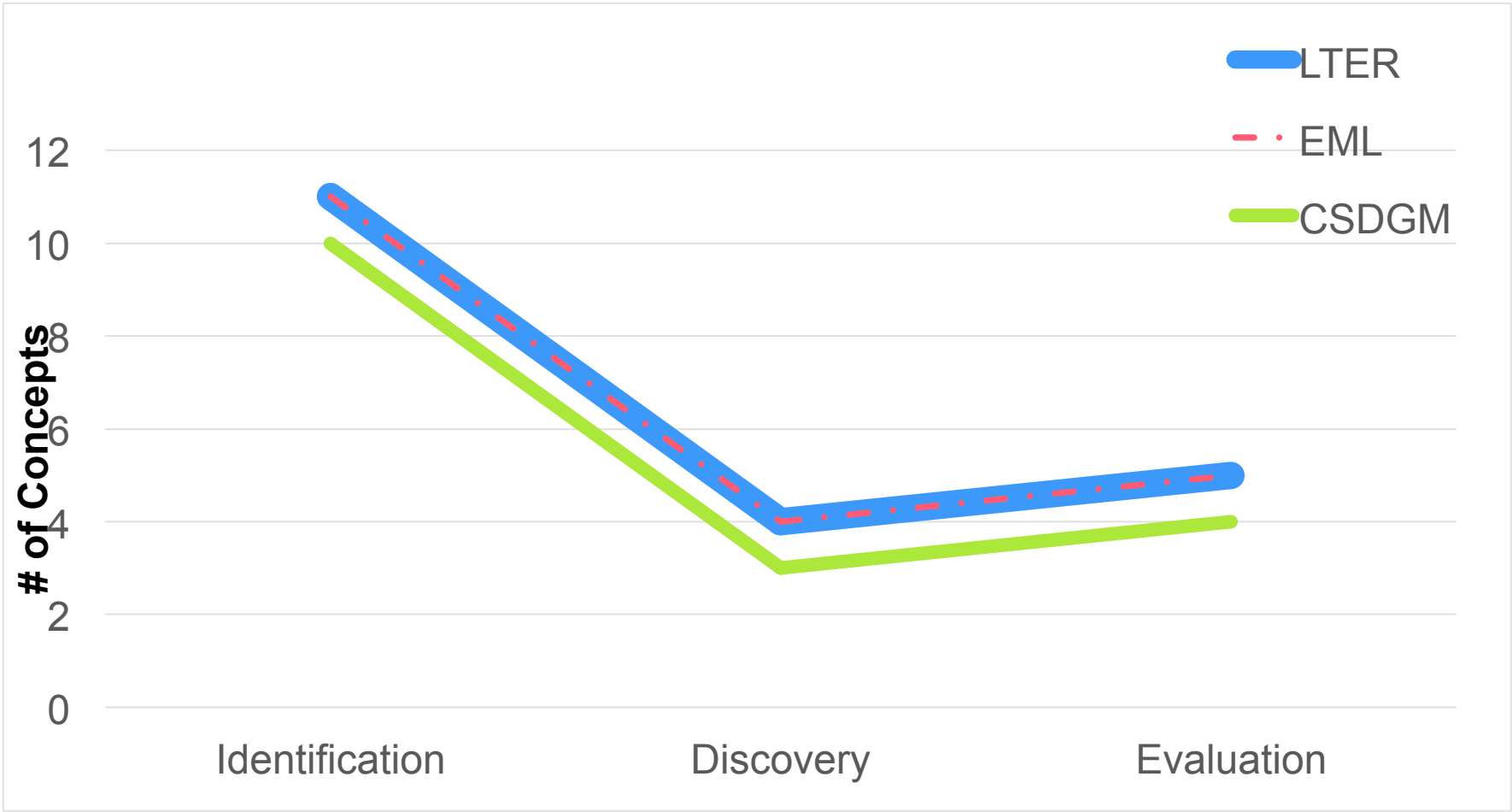


Can community developed metadata recommendations help improve metadata content within a particular community? Among different communities using the same dialect?

Improving MD across communities

EML Concepts/ Recommendation	ESA	GLEON	GOA	IOE	KNB	LTER	LTER_EU	ONEShare	TERN	TFRI	USANPN	KUBI
Resource Identifier	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Resource Title	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Author / Originator	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Metadata Contact	100%	58%	0%	0%	68%	98%	84%	0%	0%	0%	0%	0%
Contributor Name	100%	42%	95%	0%	74%	1%	0%	0%	0%	53%	100%	0%
Publisher	0%	25%	0%	0%	0%	100%	0%	94%	100%	0%	0%	0%
Publication Date	100%	50%	0%	0%	68%	100%	69%	100%	0%	0%	0%	0%
Resource Contact	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Abstract	100%	92%	100%	100%	97%	100%	88%	98%	100%	100%	100%	0%
Keyword	80%	75%	100%	96%	87%	100%	100%	100%	100%	100%	100%	100%
Resource Distribution	100%	100%	97%	100%	87%	100%	100%	94%	100%	100%	100%	100%
Taxonomic Extent	100%	0%	77%	8%	35%	0%	21%	0%	100%	12%	0%	0%
Spatial Extent	100%	92%	94%	100%	90%	100%	48%	97%	100%	100%	100%	100%
Temporal Extent	100%	92%	94%	4%	87%	100%	98%	94%	100%	35%	100%	100%
Maintenance	0%	25%	0%	0%	0%	99%	0%	0%	0%	0%	0%	0%
Resource Use Constraints	100%	92%	100%	100%	94%	99%	89%	88%	0%	82%	100%	0%
Process Step	80%	67%	94%	0%	68%	100%	100%	0%	100%	88%	100%	0%
Project Description	0%	33%	95%	8%	13%	1%	0%	94%	100%	0%	0%	0%
Entity Type Definition	0%	75%	79%	8%	16%	2%	0%	95%	0%	24%	100%	0%
Attribute Definition	0%	83%	84%	29%	23%	3%	0%	95%	0%	100%	100%	0%

LTER recommendations and CSDGM

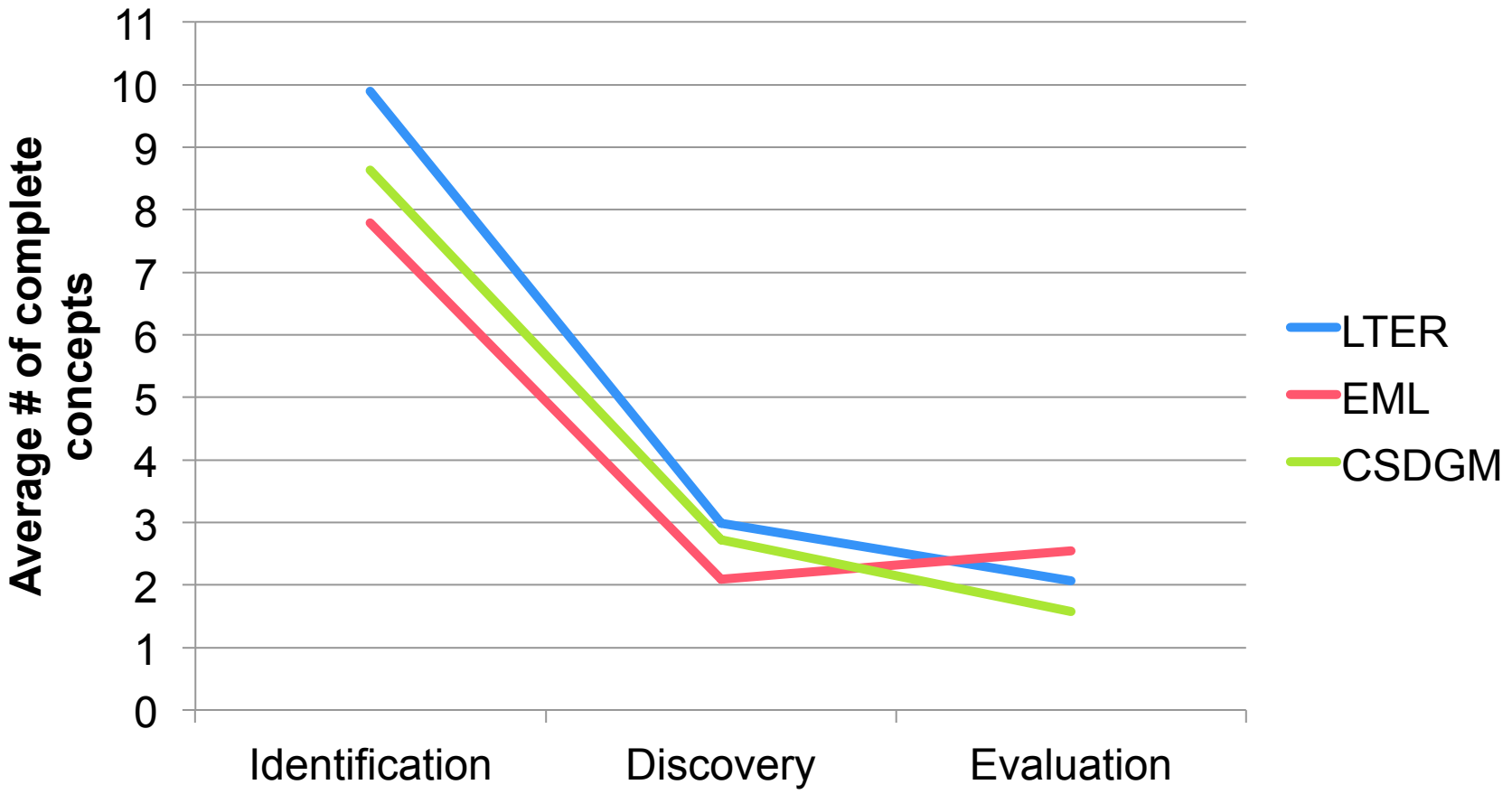


Can metadata recommendations developed in a specific dialect be used to help improve metadata in other dialects?

Recommendations across dialects

CSDGM Concepts	CDL	USGSCSAS	EDAC	EDORA	ORNLDAAC	RGD	SEAD	NMEPSCOR
Resource Identifier	-100%	-100%	-100%	-100%	-100%	-100%	-100%	-100%
Resource Title	100%	100%	100%	100%	100%	100%	100%	100%
Author / Originator	100%	100%	100%	100%	100%	100%	100%	100%
Metadata Contact	100%	100%	100%	100%	100%	100%	100%	100%
Contributor Name	100%	100%	100%	100%	100%	100%	100%	100%
Publisher	100%	26%	1%	0%	0%	0%	67%	0%
Publication Date	100%	100%	100%	0%	0%	0%	100%	100%
Resource Contact	100%	80%	100%	100%	100%	100%	67%	100%
Abstract	100%	100%	100%	100%	100%	100%	100%	100%
Keyword	100%	100%	100%	100%	100%	100%	100%	100%
Resource Distribution	0%	100%	100%	100%	100%	100%	67%	100%
Taxonomic Extent	-100%	-100%	-100%	-100%	-100%	-100%	-100%	-100%
Spatial Extent	100%	100%	100%	100%	100%	100%	100%	100%
Temporal Extent	0%	36%	95%	100%	100%	100%	89%	57%
Maintenance	100%	100%	100%	100%	100%	100%	100%	100%
Resource Use Constraints	100%	100%	100%	0%	0%	0%	100%	100%
Process Step	0%	0%	0%	0%	0%	0%	0%	0%
Project Description	-100%	-100%	-100%	-100%	-100%	-100%	-100%	-100%
Entity Type Definition	100%	98%	81%	0%	0%	0%	0%	100%
Attribute Definition	100%	98%	81%	0%	0%	0%	0%	100%

Recommendations bridging communities

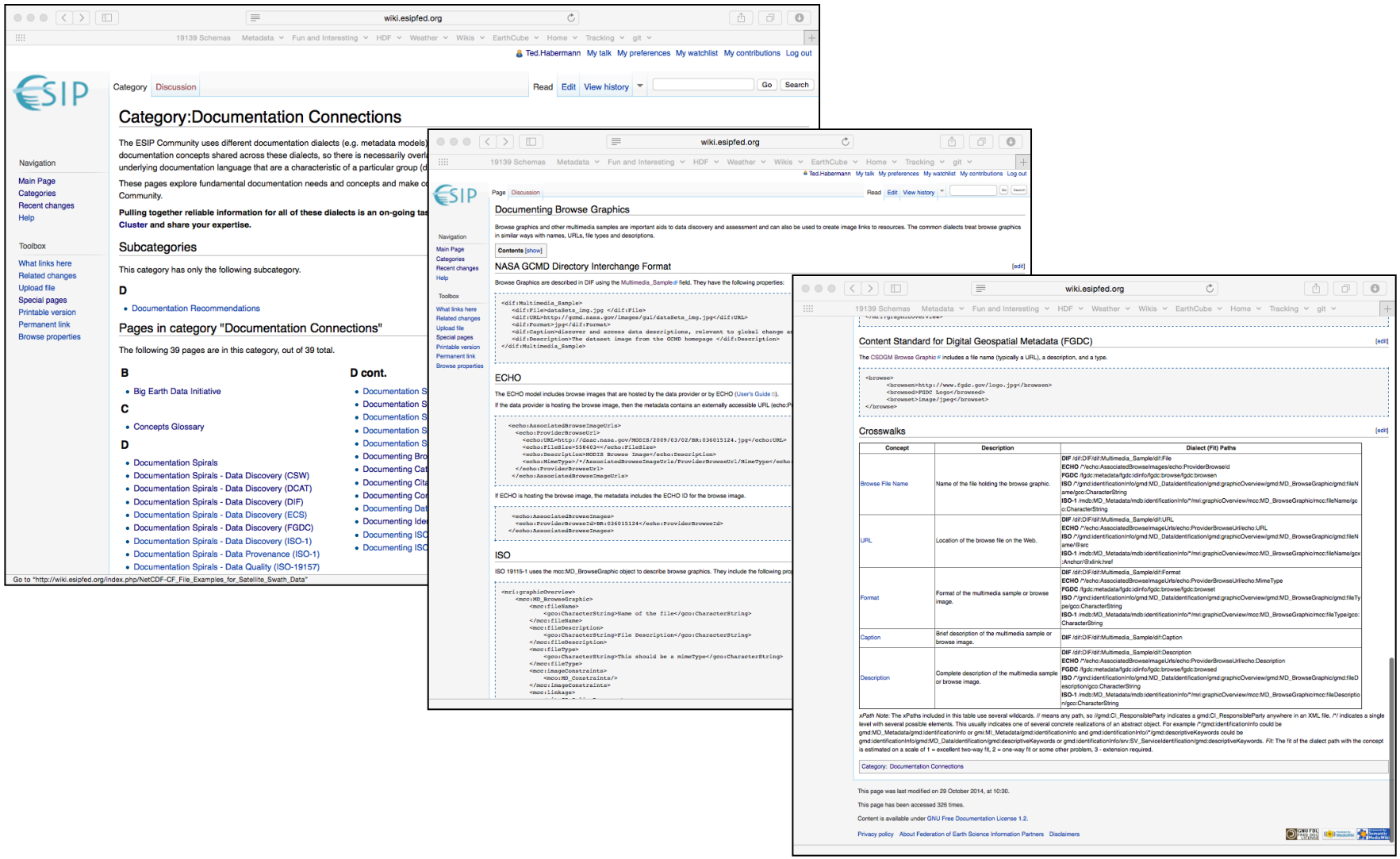


Can metadata recommendations developed in a specific dialect facilitate Communication across dialects?

Some answers…

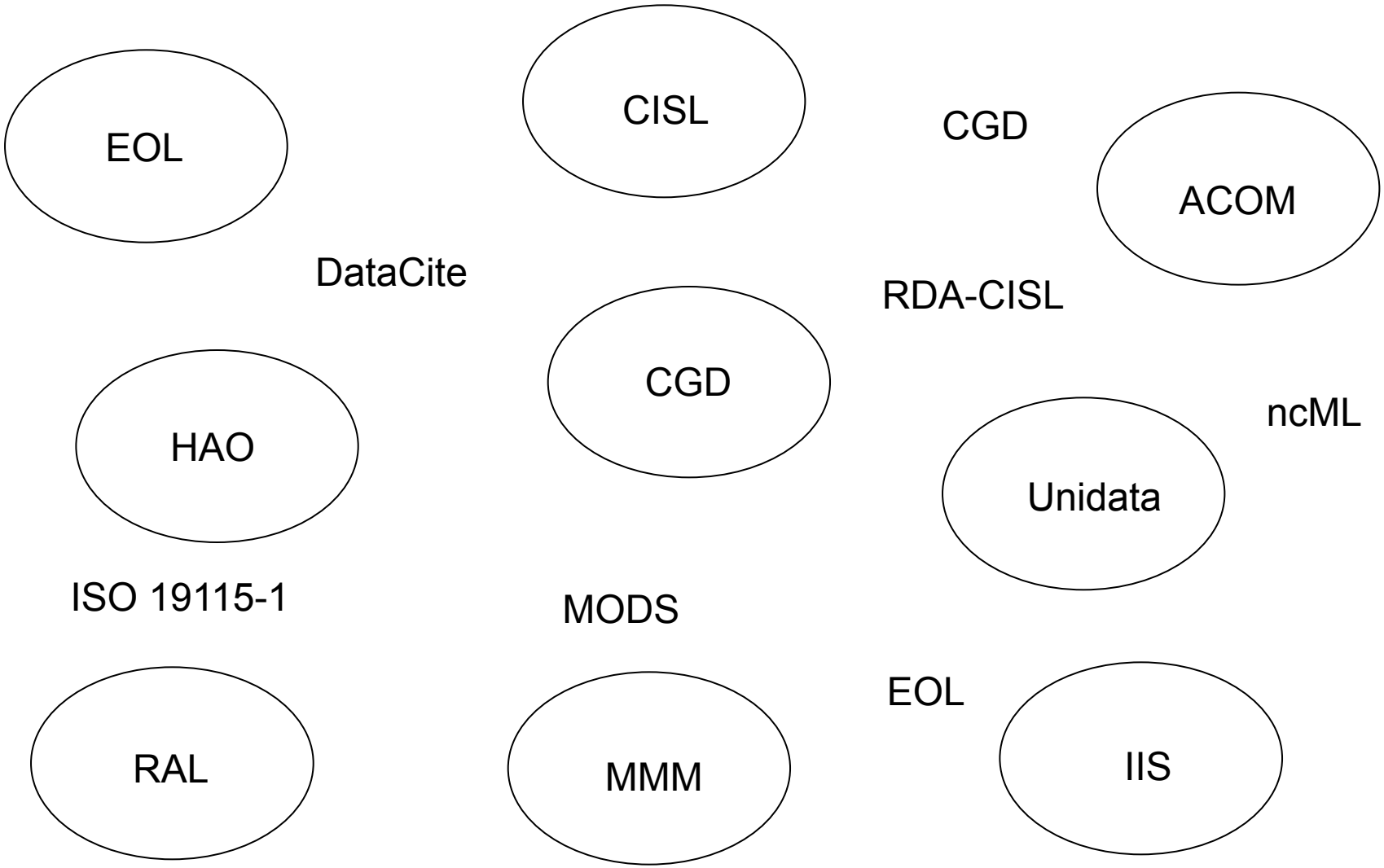
- Have the LTER recommendations improved LTER metadata completeness?
  - LTER collections are above average in complete concepts
  - Room for improvement in completeness of Evaluation concepts
  - Many LTER concepts are nearly complete, small effort to complete
- Can LTER recommendations help other communities?
  - We believe so, there seems to be a lot of alignment of MD priority concepts.
  - What strategies might be useful to communities to help improve completeness?
- Can metadata recommendations developed in a specific dialect be used to help improve metadata in other dialects?
  - CSDGM dialect contains most of the concepts found in the LTER recommendation spirals, and therefor these recommendations can be easily applied to CSDGM collections
  - CSDGM collections are complete with respect to most of the LTER recommended concepts

Guidance Documentation



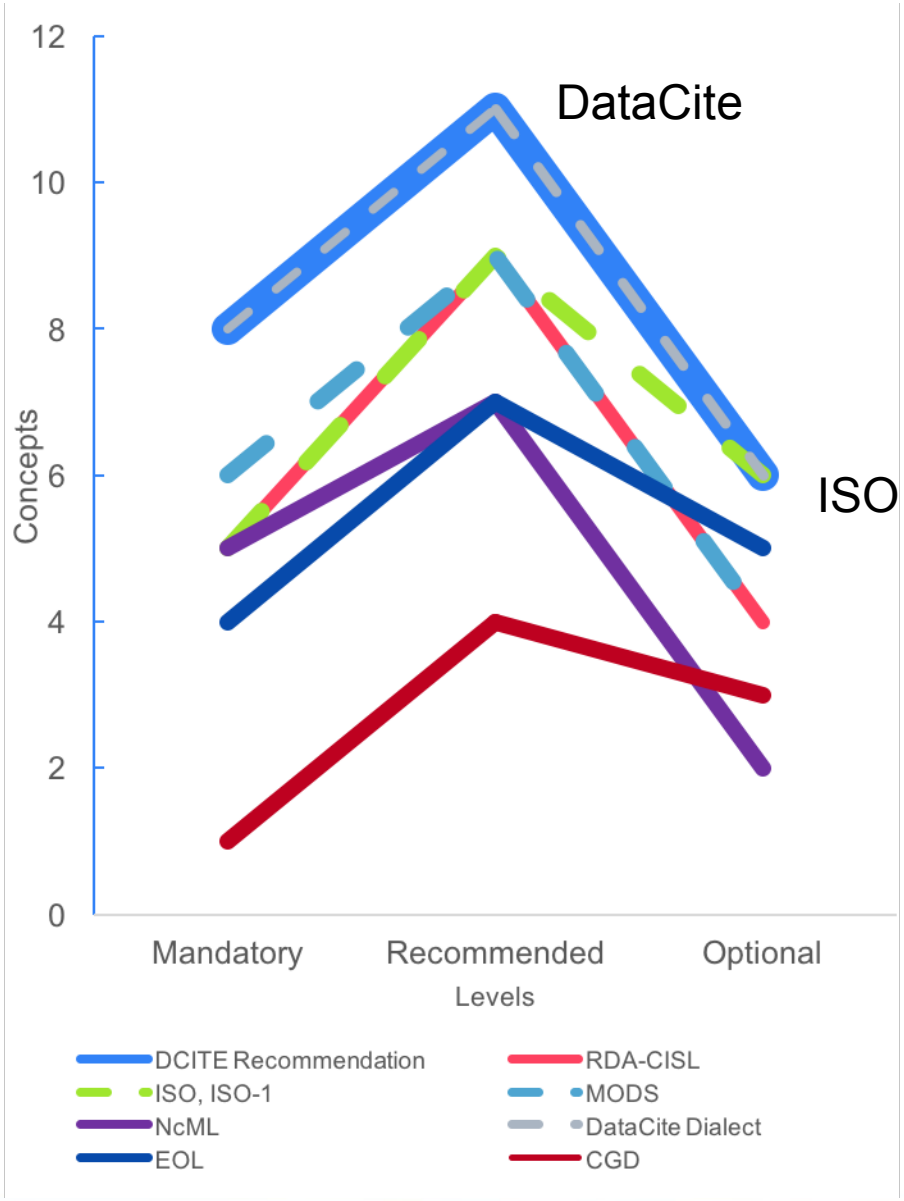
http://wiki.esipfed.org/index.php/Category:Documentation_Connections

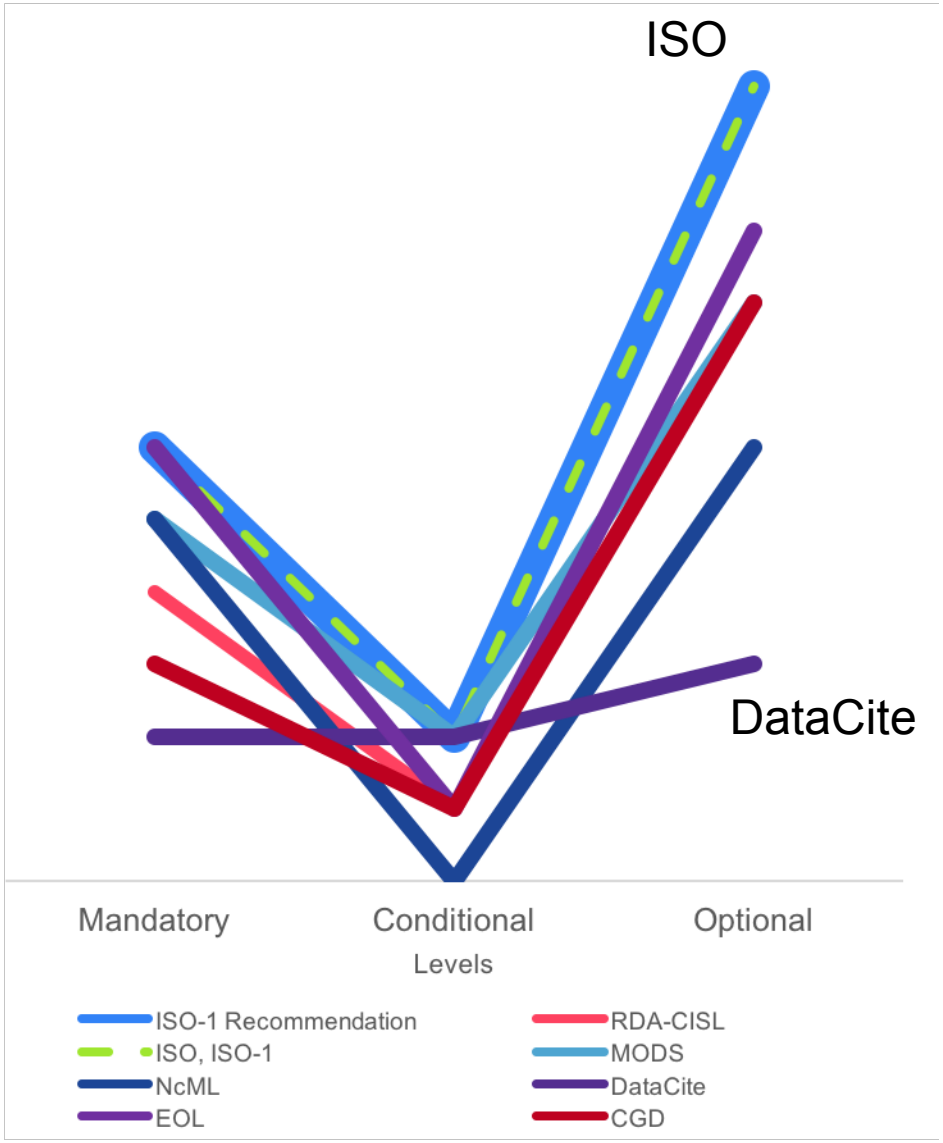
The UCAR Labs and Dialects



Recommendations Comparison

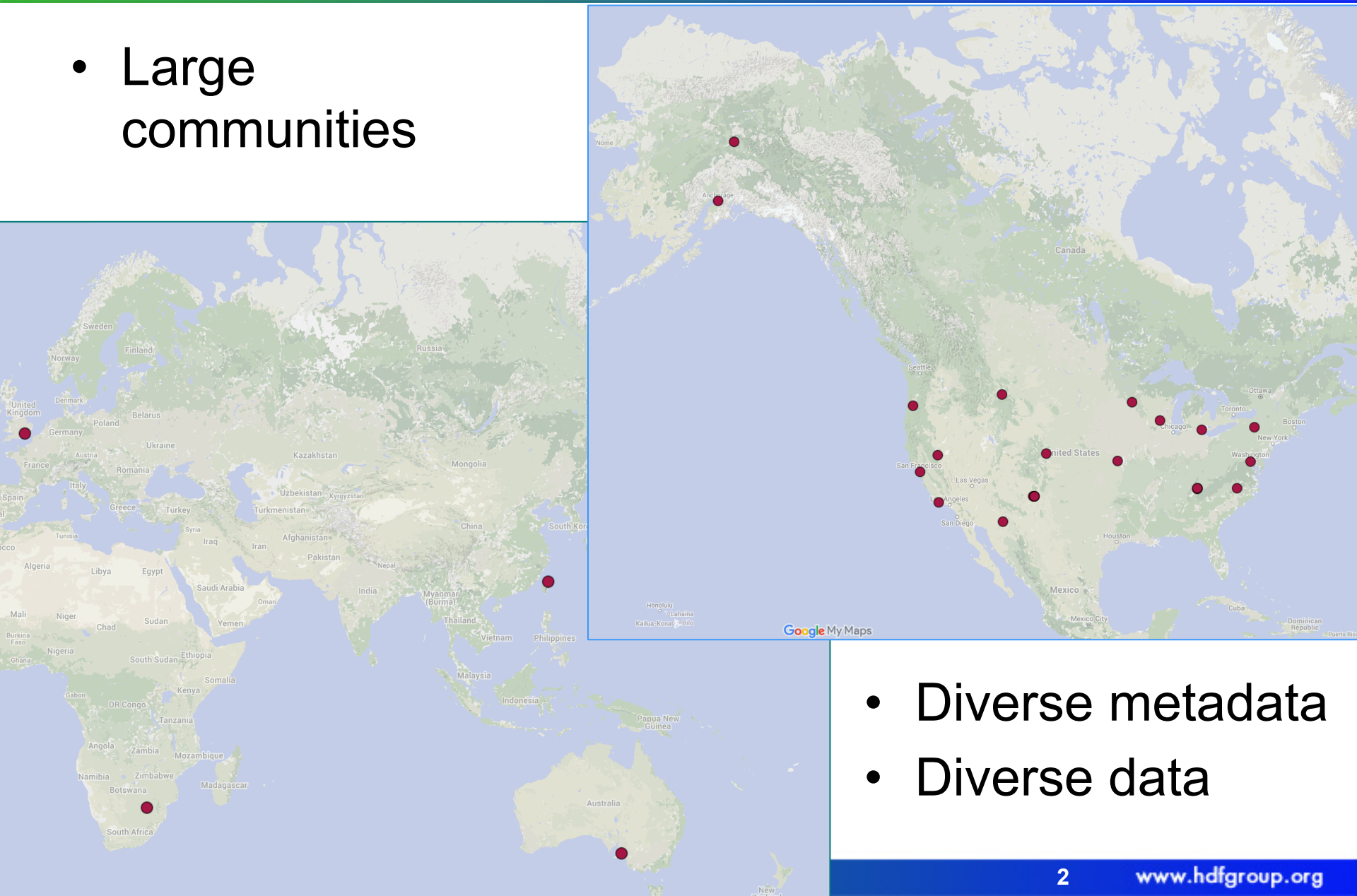
What recommendation fits our science?





DataONE Repositories

- Large communities



- Diverse metadata
- Diverse data

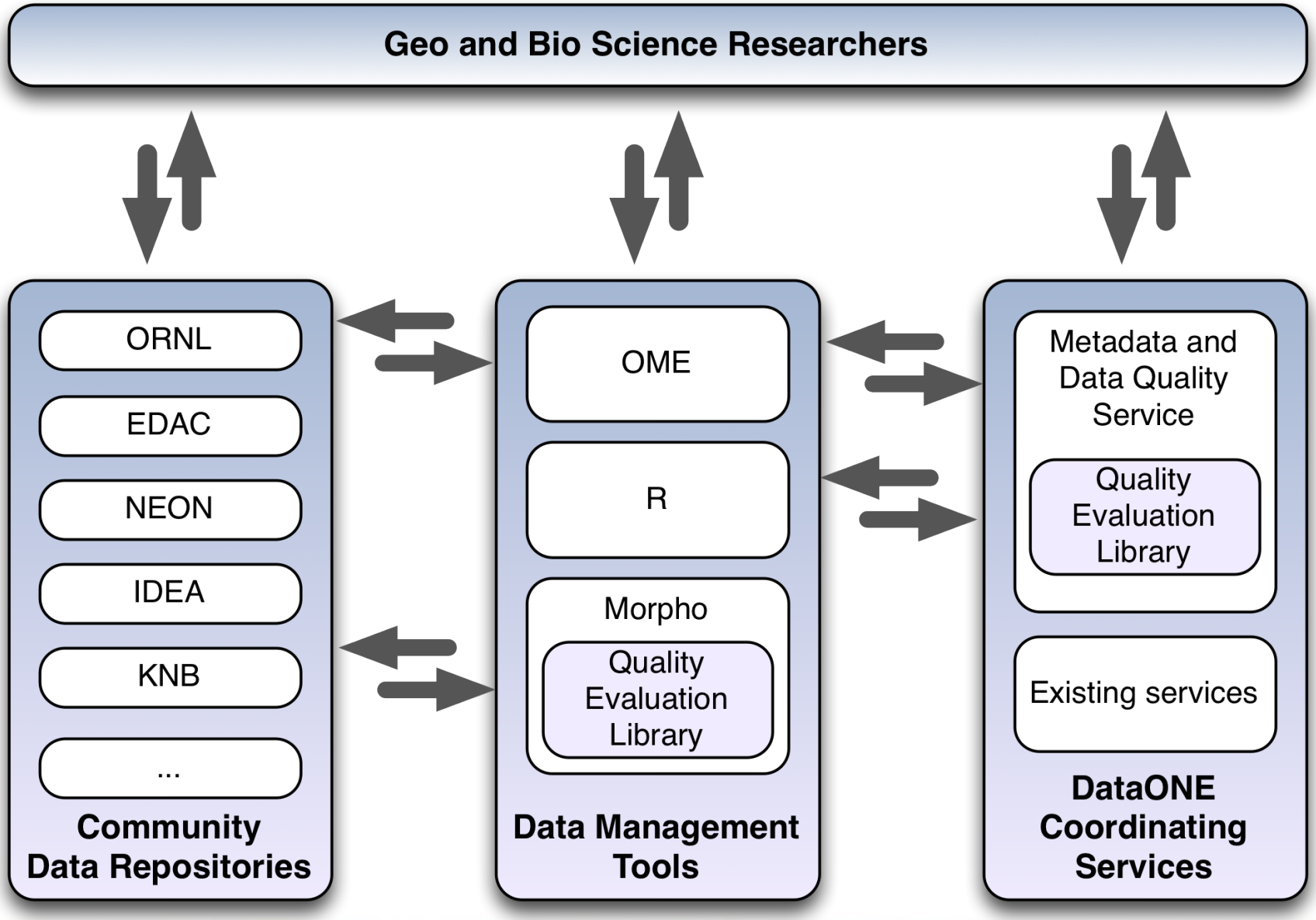
MetaDIG Tools and services

- 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



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
…			

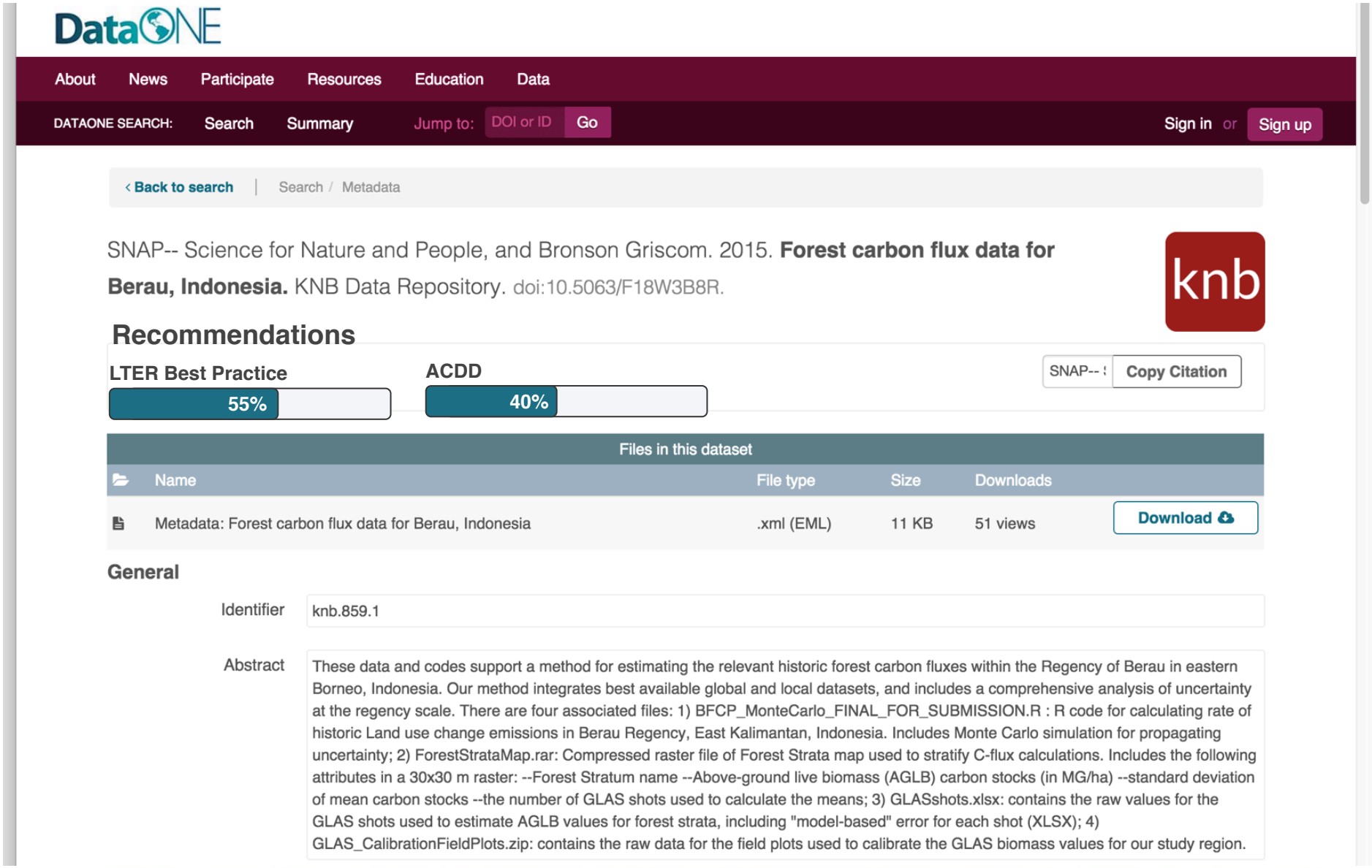
- Checks in Java, R, Python
- Categorized by function (discovery, re-use, …)
- Operate across dialects (EML, CSDGM, ISO19139)

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



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

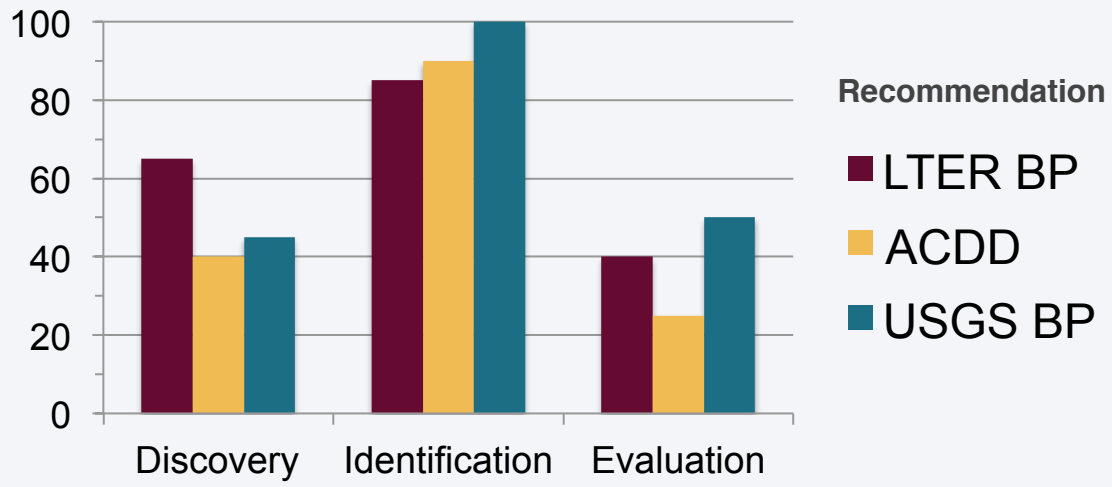
knb Gregory Goldsmith. 2016. **Data from: Plant-O-Matic: A dynamic and mobile guide to all plants of the Americas.** KNB Data Repository. knb.909.8.

3

knb 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.