Project 418 Report

Douglas Fils Eric Lingerfelt Adam Shepherd

~261 days ~430 commits ~1.65 commits / day

EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

BALTO
CHORDS
CDF RWG

MagIC

BCO-DMO
Biological & Chemical Oceanography Data Management Office

Neotoma

HYDROSHARE

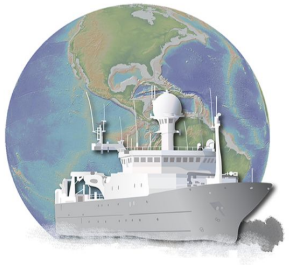
UNAVCO

IEDA
INTERDISCIPLINARY
EARTH DATA ALLIANCE

OpenTopography
High-Resolution Topography Data and Tools

LinkedEarth

IRIS



R2R

CSDCO
Continental Scientific Drilling
Coordination Office

Open
Core

JOIDES Resolution
Science Operator

UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

NSF

Project 418 The Basics

- Publish via web architecture
- Schema.org and community vocabularies to provide context
- Consume (subscribe) published data through web architecture
- Indexes & services

A Full stack prototype to explore the potential

Use the Web

Project 418

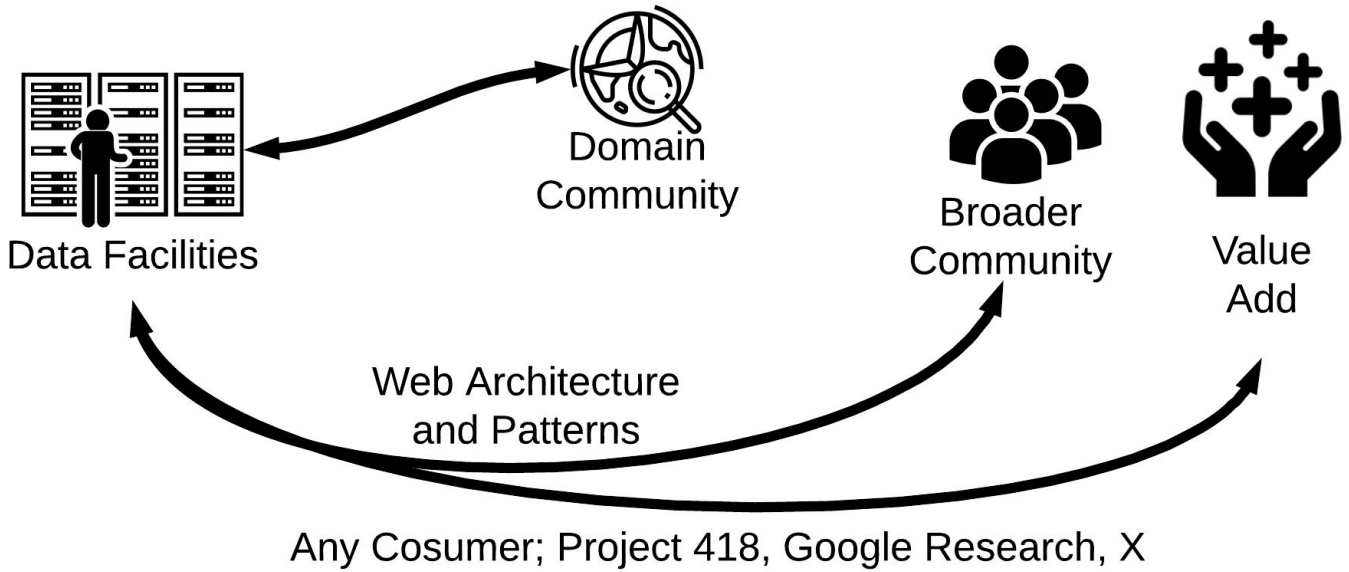
EarthCube
TRANSFORMING GEOSCIENCES RESEARCH

UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

NSF

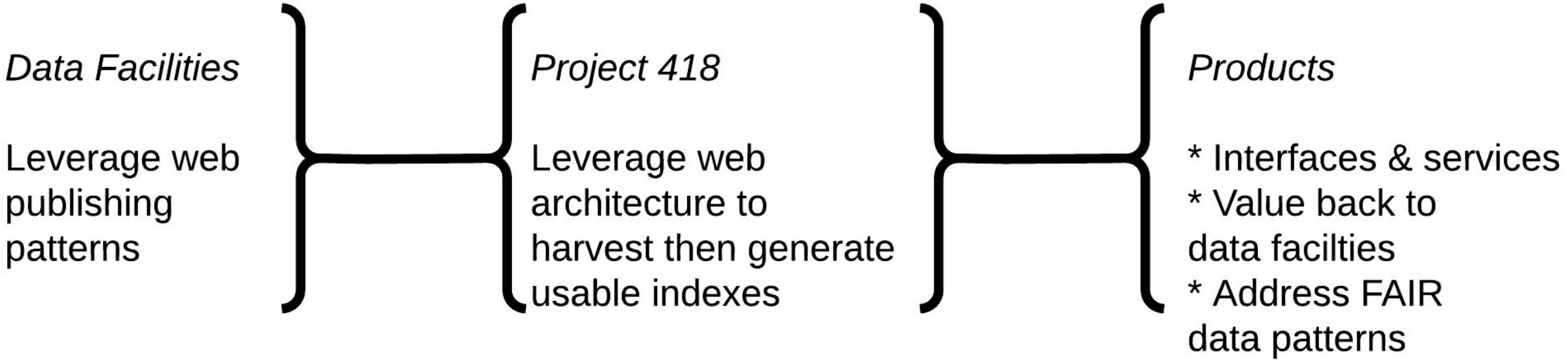
High Level View

Reach broader community & feed value back to facility

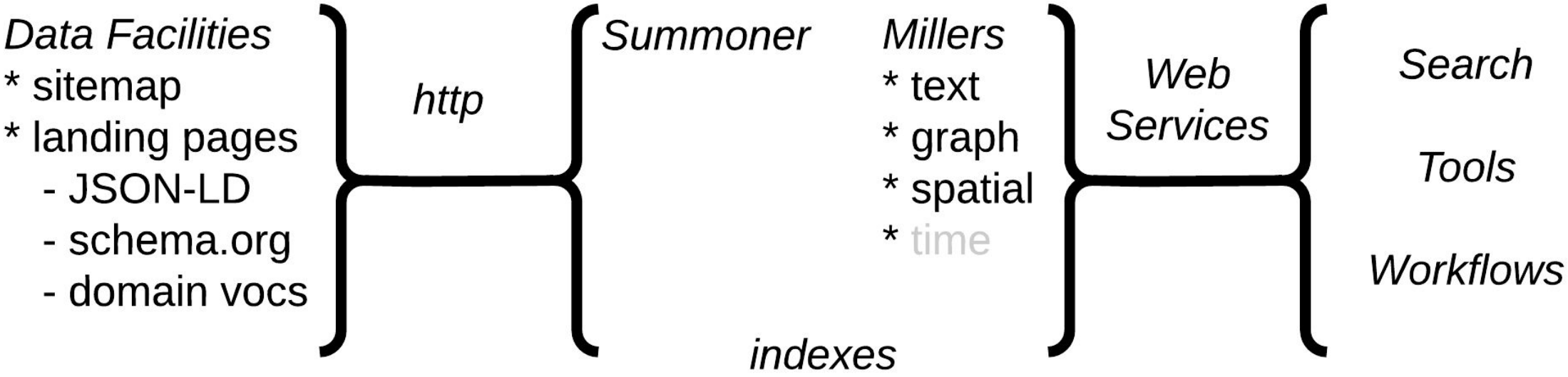


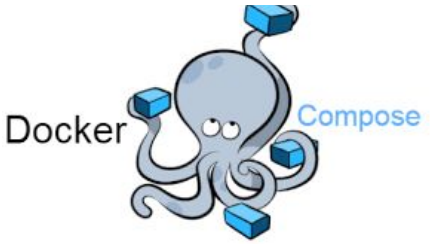
Players with their actions

Players again, with actions



Players… with a bit more detail….





P418: Implementation (vs Philosophy)

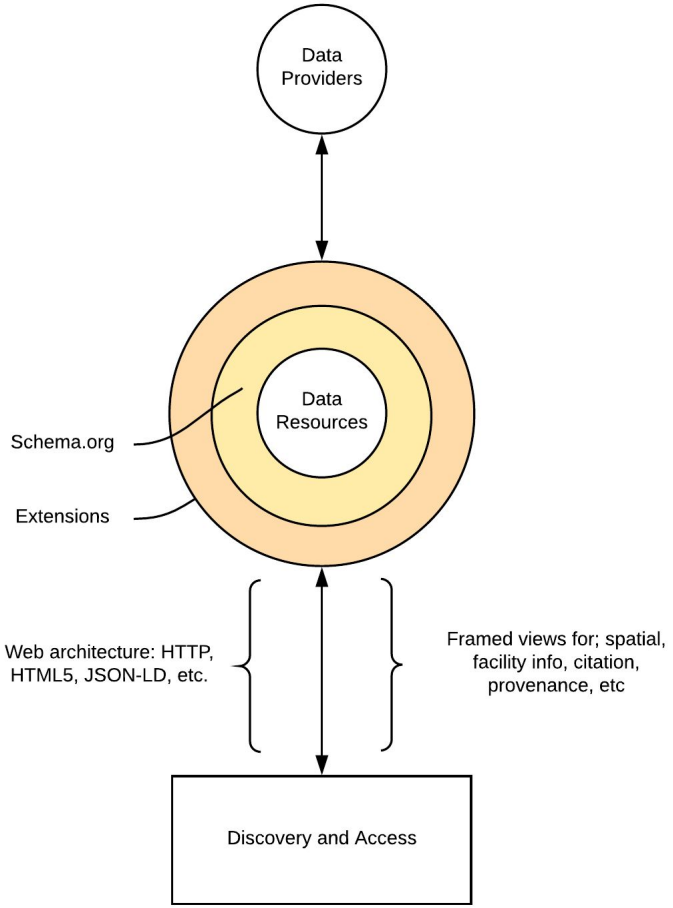
SCOPE: Working with a set of NSF data facilities to *demonstrate publishing approaches* for schema.org/Dataset and extensions (special thanks to them!)

Use schema.org as a base vocabulary with extensions to connect metadata publishing workflows to community vocabularies.

SCOPE: *Implement harvesting and interface packages* to further explore the full pipeline and provide feedback.

https://geodex.org (an implementation of Project 418)

- API listing
- UI Gallery
- Links to docs, vocs and code



Data In Context

Project 418

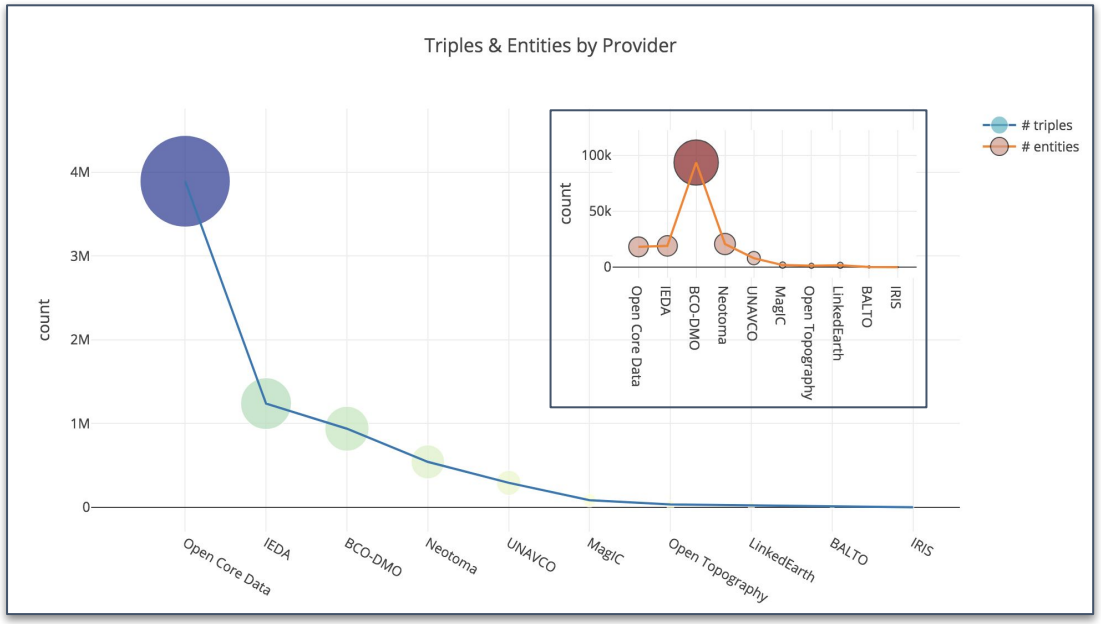
EarthCube
TRANSFORMING GEOSCIENCES RESEARCH

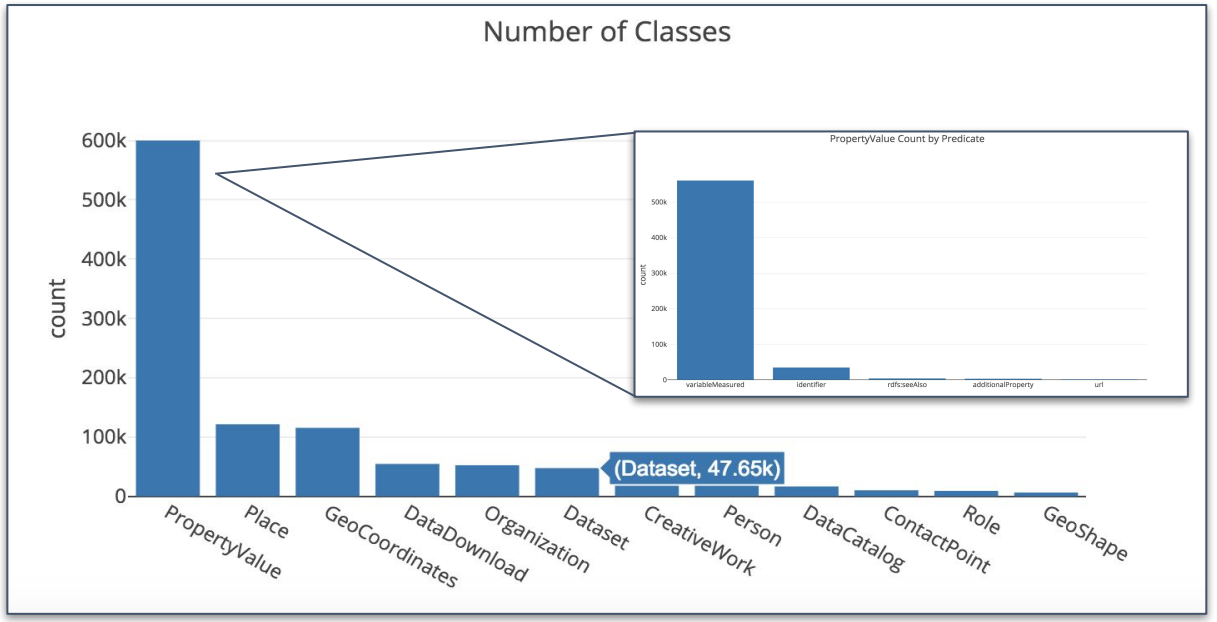
UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

NSF

Summary Statistics

1,113,210 entities
7,087,380 triples





47,650 Dataset
54,665 DataDownload
599,960 PropertyValue
~ 35k Identifiers
~560k Dataset Variables

EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

Vocabulary Use - Google Recommended

Dataset Properties	Google Requires / Recommends	Provider Usage	Dataset Coverage	
			Implemented	*Overall*
@context	**Required**. Set `@context` to "http://schema.org/"	**80%**	omitted ending slash: 'http://schema.org'	
@type	**Required**. Set `@type` to "Dataset"	**100%**	**47,650 datasets**	*n/a*
name	**Required**. A descriptive name	**80%**	99.9%	73%
description	**Required**. A short summary	**70%**	97%	69%
url	*Recommended*.	**70%**	100%	62%
citation	*Recommended.*	**60%**	100%	36%
keywords	*Recommended.*	**70%**	99.9%	66%
spatialCoverage	*Recommended.*	**80%**	92%	91%
temporalCoverage	*Recommended.*	**10%**	15%	<1%
variableMeasured	*Recommended.*	**30%**	83%	40%
version	*Recommended.*	**40%**	95%	25%
sameAs	*Recommended.* Same data, different URL.	**10%**	100%	<1%

EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

https://developers.google.com/search/docs/data-types/dataset

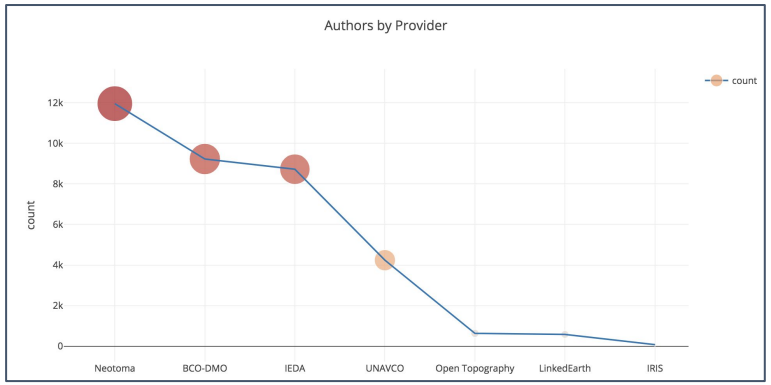
80 - 100%	50 - 79%	0 - 49%

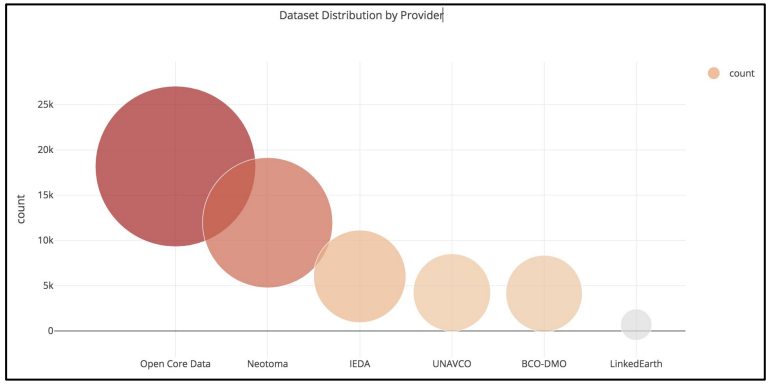
UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

NSF

Vocabulary Use - P418 Recommended

Dataset Properties	Provider Usage		Dataset Coverage	
			Implemented	*Overall*
identifier	**30%**	10,556 datasets	100%	22%
author/creator/contributor	**80%**	28,765 datasets	98%	69%
funder *(not awards)*	**30%**	4,069 datasets	78%	9%
distribution	**60%**	45,221 datasets	100%	95%
license	**70%**	42,523 datasets	98%	89%
hasPart *ex: linking PhysicalSamples to Datasets*	**10%**	122 datasets	2%	<1%





"*What about Data APIs?*"

- **3 providers**: Search endpoints, SWAGGER, SPARQL, VoID, OGC CSW

80 - 100%	50 - 79%	0 - 49%

EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

github: [earthcubearchitecture-project418/p418Vocabulary]

UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

NSF

Vocabulary Use - External Vocabularies

- Some providers used external vocabularies
  - EarthCube Building Blocks - EarthCollab & GeoLink
  - Datacite Ontology - DOIs and ORCID
  - ViVO Ontology - Datasets

Opportunity to improve search precision

- Geoscience Standard Names,
- SWEET Ontologies,
- GCMD Keywords,
- etc.

github: earthcubearchitecture-project418/assay-data

Principles Over Project

Philosophy over Implementation

EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

NSF

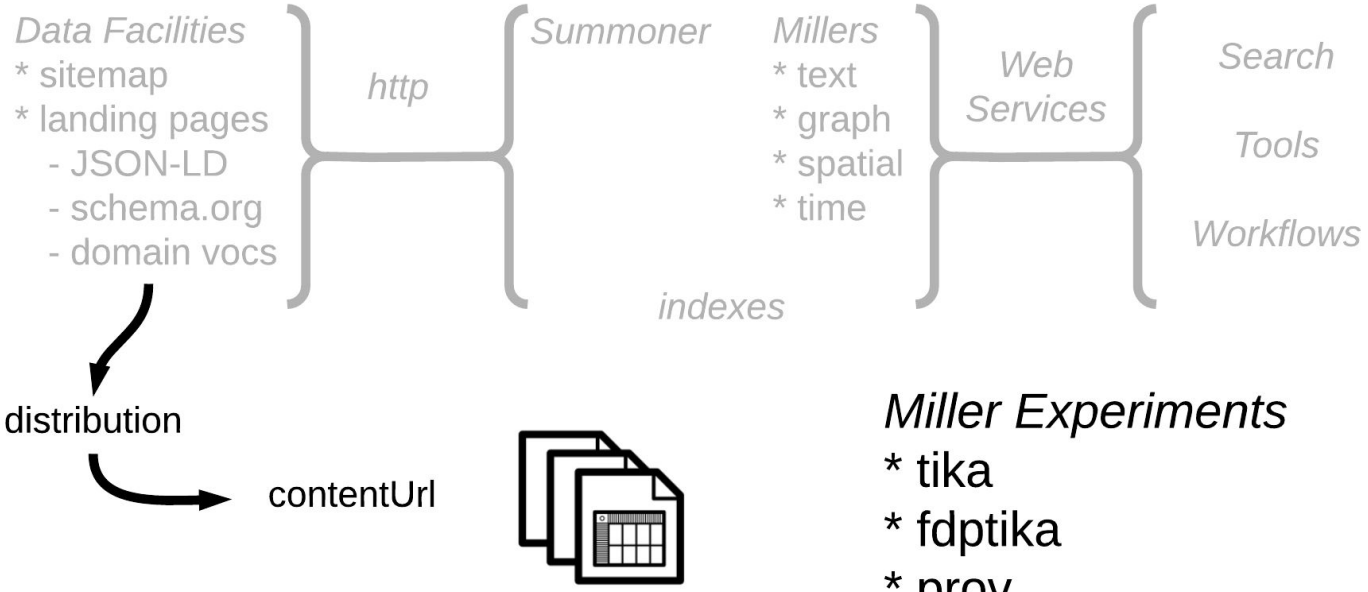
Project 418: It does more than just indexes!

Once you have a simple architecture that gateways machine readable resources

There is a LOT you can do!

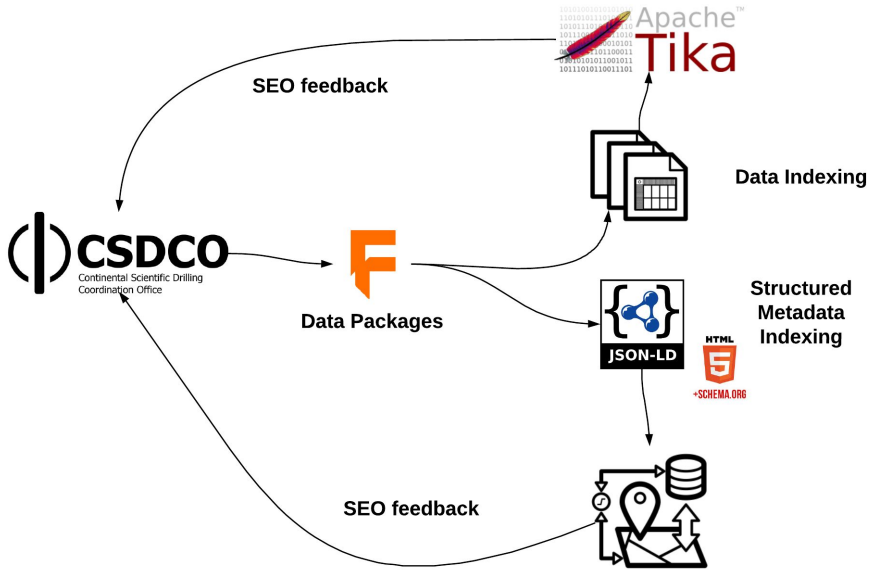
Pervasive Data Mobility

Leverage the structured metadata to identify and act on distribution links



Miller Experiments
* tika
* fdptika
* prov
* logs (from services, not a miller)

- Shows the need for PIDs
- Prov to address “implicate triple” issue



EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

NSF

Aggregators Dilemma

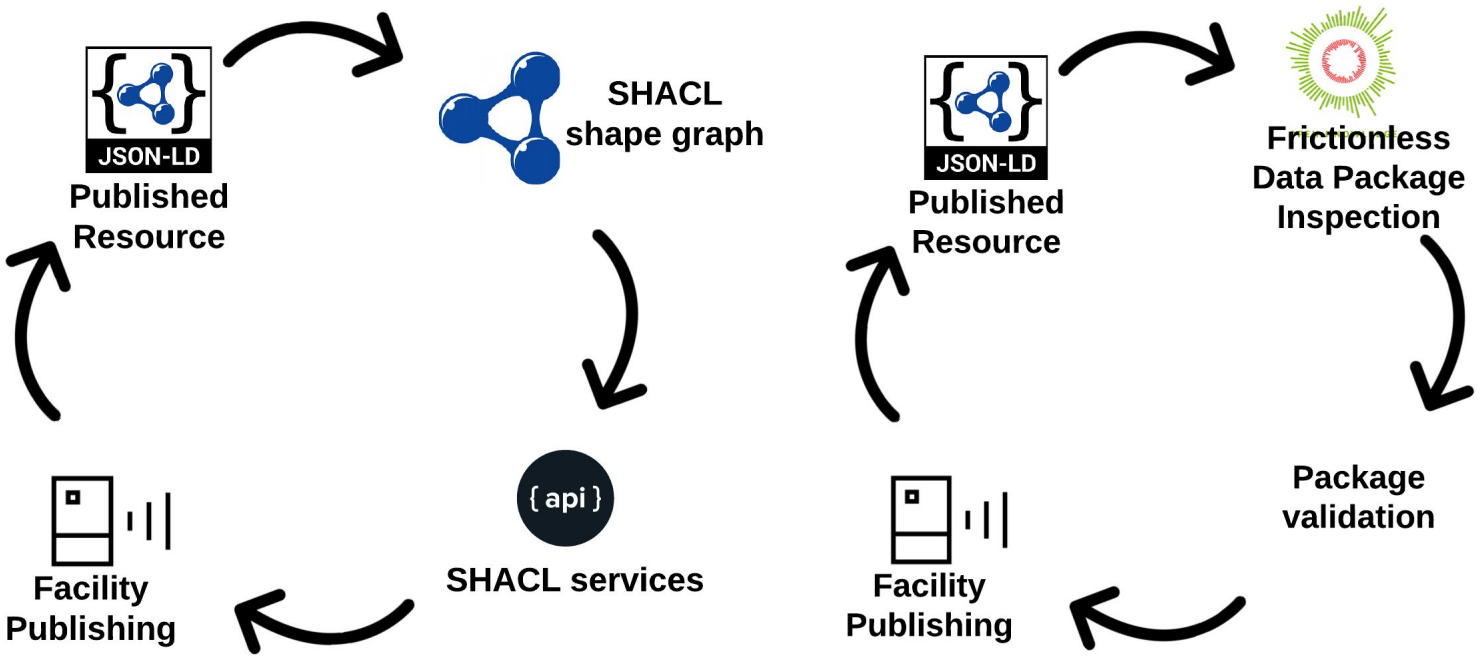
- Aggregate to server a larger community than each individual provider
- Enage a broader community while *not disintermediating existing community* relationships of these authoritative providers
- Aggregation should be a Value Add Service (VAS). The “cost” of being aggregated must be less that benefits
  - drive traffic (likely not enough)
  - provide assessment (alignment to community expectations)
  - provide a means to form and/or engage new communities
- Explicate vs Implicate triples (drove us to prov)

PROV: A quick note

- JSON-LD encoding can easily lead to a need for implicate triples
  - We are trying to avoid adding any data
- Prov allows us to connect resources to the providers via organizational information (avoid generating implicate triples)
- A working point with the facilities to potentially remove this?
- The PROV graph does offer some information back to the facilities and connect log info back
- Potential point of interaction with the ESIP Lab's PROV work

Value Add Services examples (Data SEO)

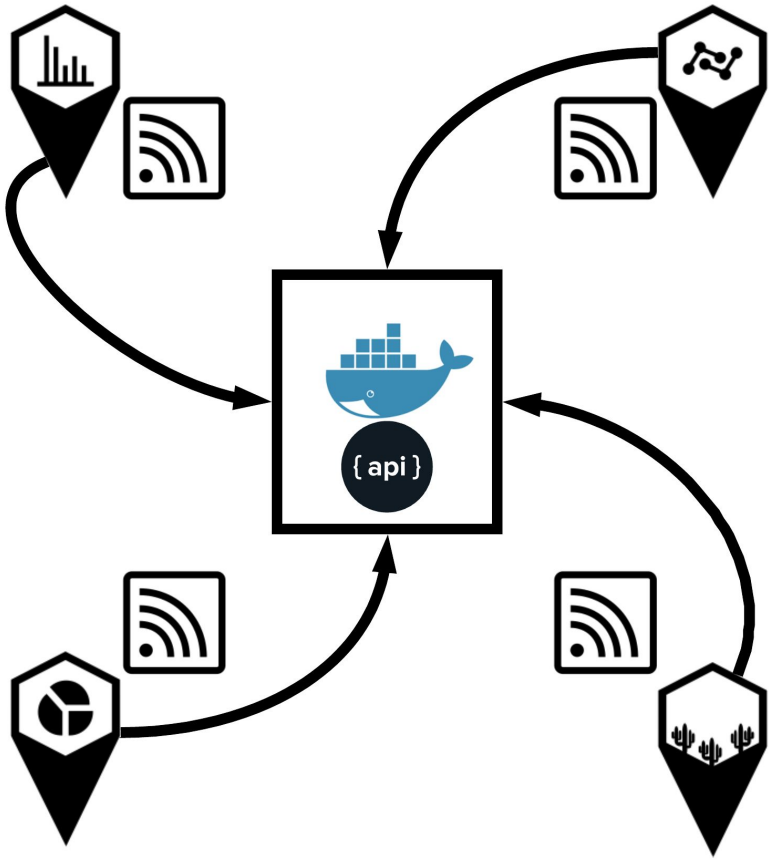
"Millers" for shape testing and package inspection (more than just indexes)



- Shape graphs can be defined by a community
- Engages the science, publishing and data facility communities
- Addresses a process issue exposed in P418

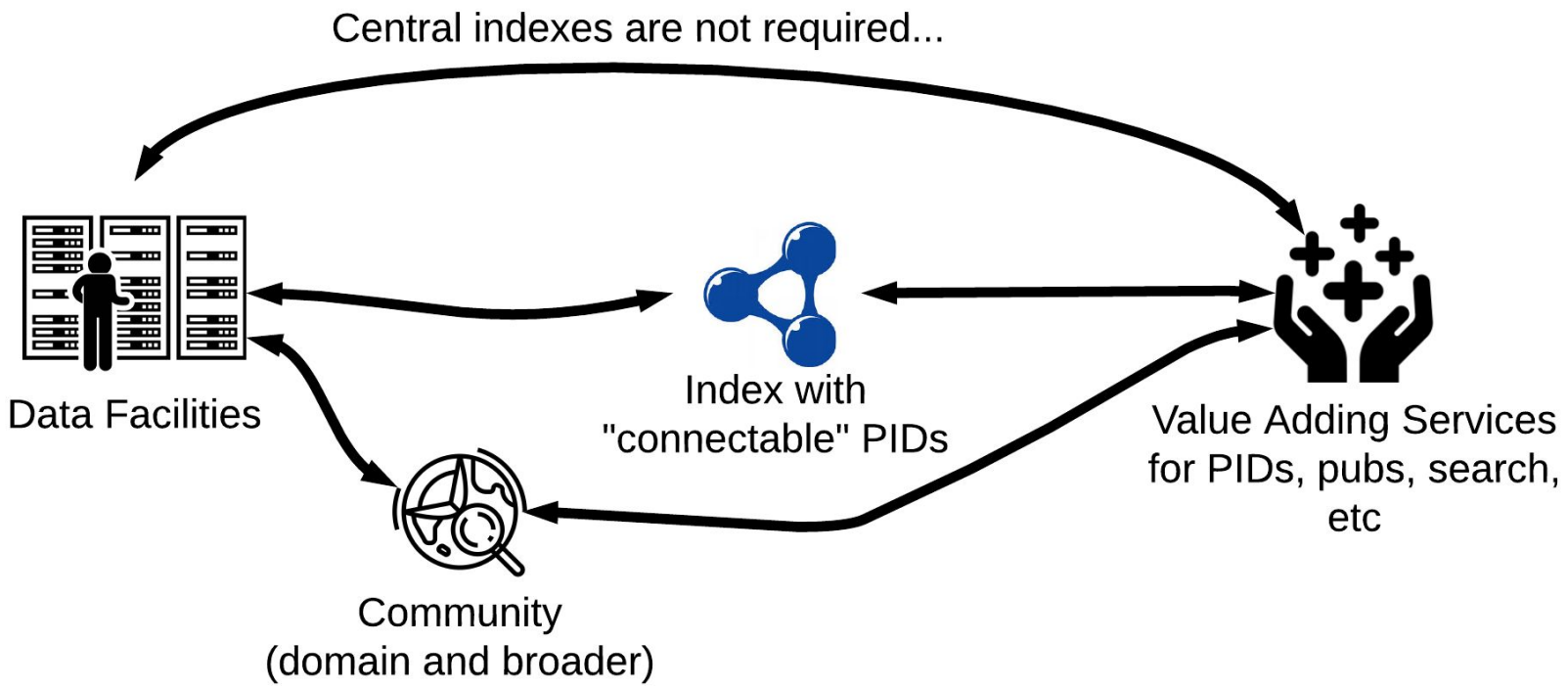
UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

µP418 why scale up.. when you can scale down



- full P418 stack runs on my laptop
- raises question of distributed interop (RDA is on this too)
- P418 could be used by domains of practice to build shared indexes to enhance and align a particular community experience

Aggregators, Platforms and Networks



If we enable N indexes, we want to think now about sharing connections and affordances of indexed resources

A Process and a Philosophy, not a Product

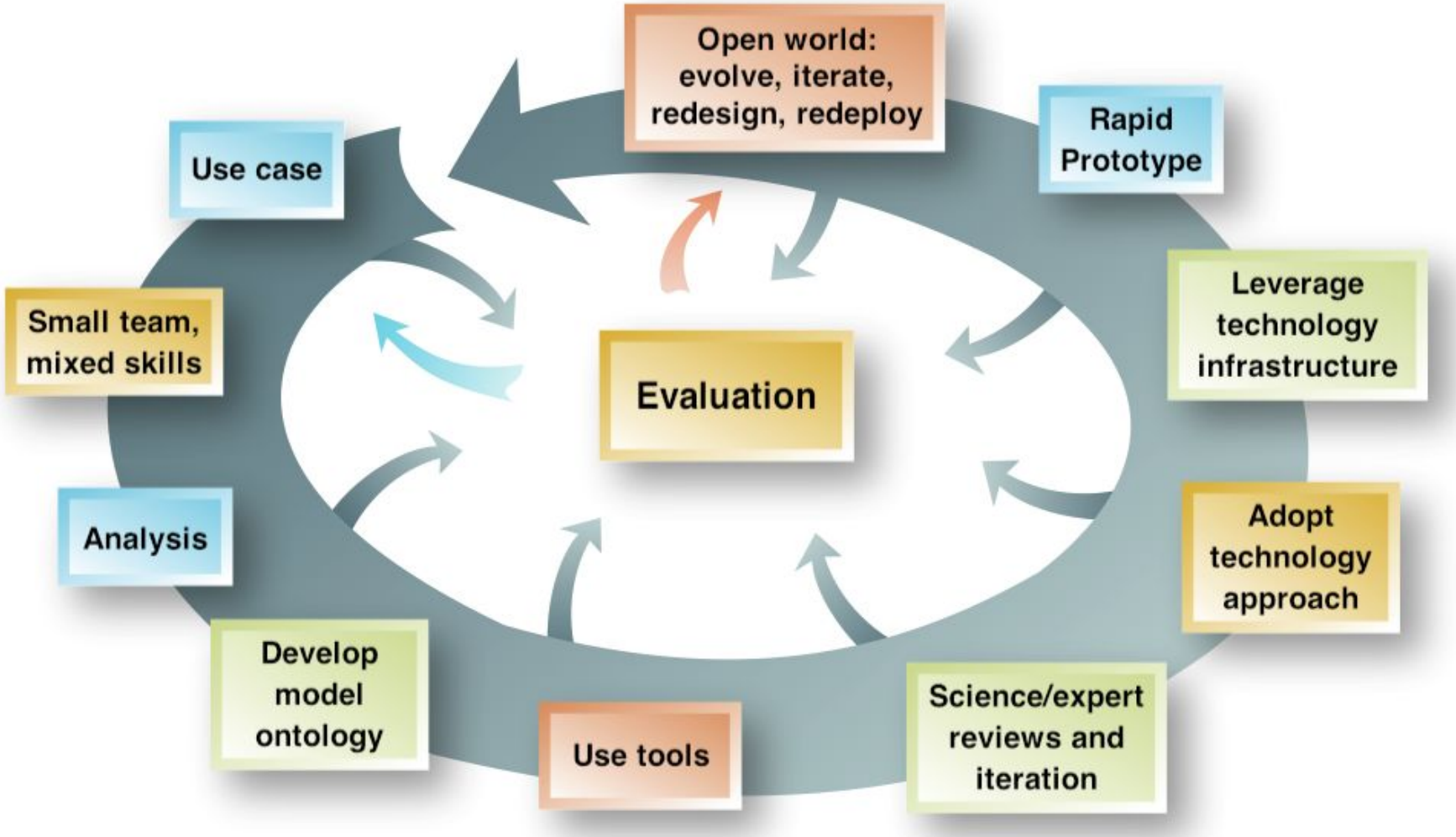


Image credit: Peter Fox https://tw.rpi.edu/web/doc/TWC_SemanticWebMethodology

- I've been seeing Peter's image for years… I've come to really appreciate it in this project

UCAR UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

Use The Web
Data In Content
Principle Over Project

Project 418
https://www.earthcube.org/group/project-418

https://github.com/earthcubearchitecture-project418
https://geodex.org

EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

BALTO
CHORDS
CDF RWG

MagIC

BCO-DMO
Biological & Chemical Oceanography Data Management Office

Neotoma

HYDROSHARE

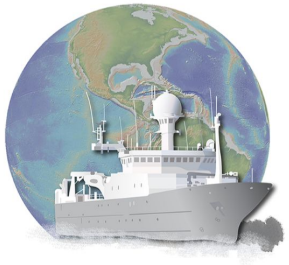
UNAVCO

IEDA
INTERDISCIPLINARY
EARTH DATA ALLIANCE

OpenTopography
High-Resolution Topography Data and Tools

LinkedEarth

IRIS



R2R

CSDCO
Continental Scientific Drilling
Coordination Office

Open
Core

JOIDES Resolution
Science Operator

UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

NSF

Thanks Also To

ESSO

XSEDE
Extreme Science and Engineering Discovery Environment

Research at Google

Project 418 Project Advisory Team

Rick Benson - IRIS
Fran Boler - UNAVCO
Matt Mayernick - NCAR
Tanu Malik - DePaul University
Sarah Stamps - Virginia Tech
Steve Kuehn - Concord University

ESIP

RDA
RESEARCH DATA ALLIANCE

EarthCube Community

UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

NSF

Thanks