Data Federation via Metadata

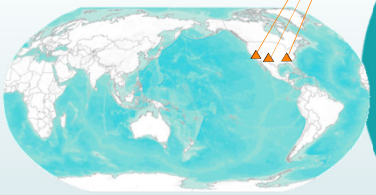
Bill Michener
University of New Mexico

EUDAT 2nd Conference
Rome, Italy
October 29, 2013

DataONE
UNM COLLEGE of UNIVERSITY LIBRARIES & LEARNING SCIENCES

Technology

Three major components for a flexible, scalable, sustainable network

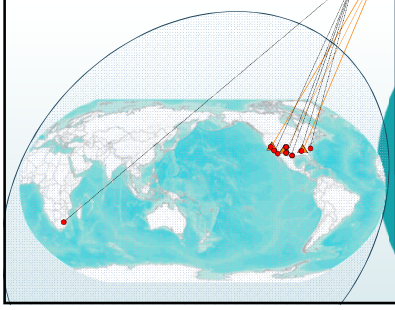


Coordinating Nodes
- retain complete metadata catalog
- indexing for search
- network-wide services
- ensure content availability (preservation)
- replication services

THE UNIVERSITY of NEW MEXICO · OAK RIDGE · NCEAS · UCSB · UT

Technology

Three major components for a flexible, scalable, sustainable network



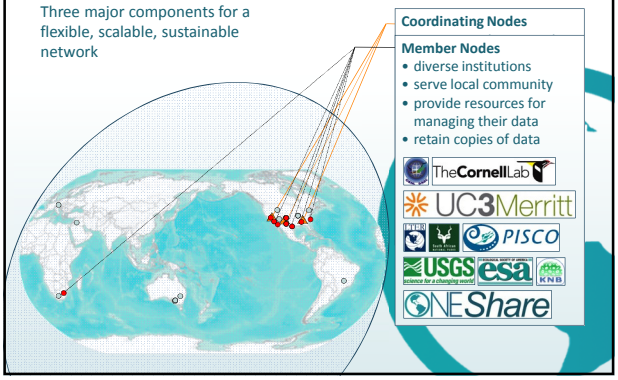
Coordinating Nodes

Member Nodes
- diverse institutions
- serve local community
- provide resources for managing their data
- retain copies of data

TheCornellLab · UC3 Merritt · PISCO · USGS · esa · KNB · ONEShare

Technology

Three major components for a flexible, scalable, sustainable network



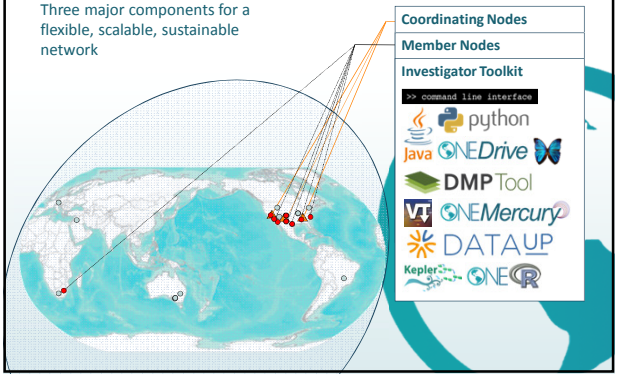
Coordinating Nodes

Member Nodes
- diverse institutions
- serve local community
- provide resources for managing their data
- retain copies of data

TheCornellLab
UC3 Merritt
PISCO
USGS esa KNB
ONEShare

Technology

Three major components for a flexible, scalable, sustainable network



Coordinating Nodes

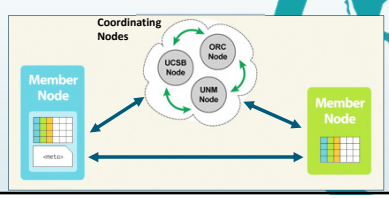
Member Nodes

Investigator Toolkit

>> command line interface
python
Java ONEDrive
DMP Tool
ONEMercury
DATAUP
Kepler ONE R

Preserve Data and Metadata

- Metadata mirrored at Coordinating Nodes
- Data replicated between Member Nodes
- CNs manage copies
- Checksums recorded and verified
- Promote quality metadata

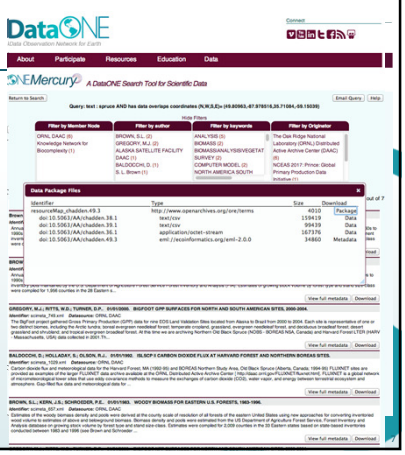


Discovery

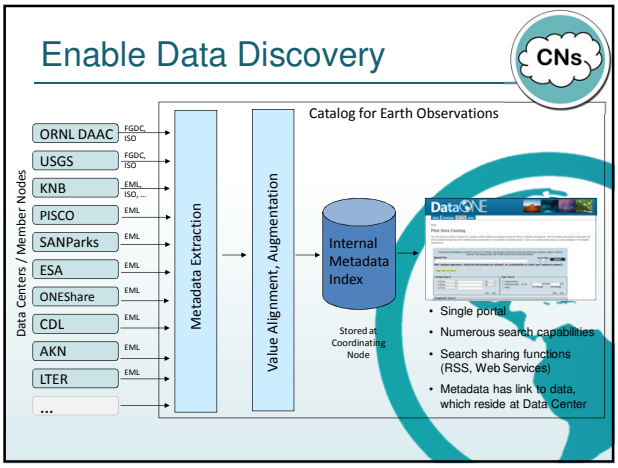
State:
- Diversity of participating nodes and resources
- Coordinating Nodes collate and index content

Issues:
- Many disciplines
- Legacy content
- Heterogeneity

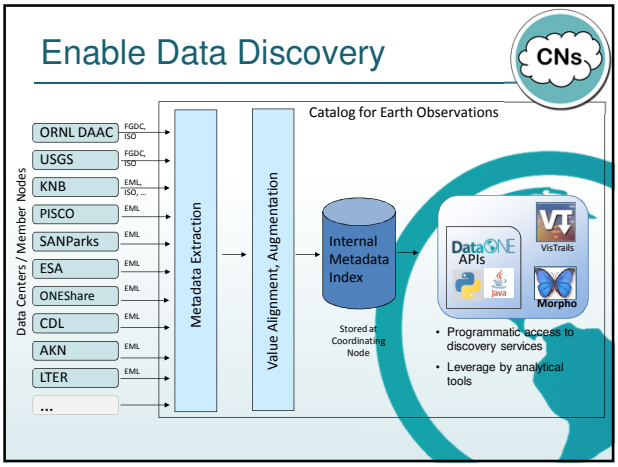
Topics:
- Precision
- Recall
- Data Integration



Enable Data Discovery



Enable Data Discovery



Metadata Quality

- Data discovery
  - Title, abstract, investigators, etc.
  - Data extent (over space, time, and taxa)
  - Associated research papers
- Data interpretation and integration
  - Methodological details allow re-use
  - Semantics allow improved integration
- Analysis and modeling
  - Automate mechanical processing
  - Allows focus on science issues not plumbing

Metadata Formats

- Extract and index common fields from metadata standards
  - Ecological Metadata Language (EML)
  - FGDC Biological Data Profile (BDP)
  - ISO 19115 Geospatial Metadata
  - Dublin Core
  - Darwin Core
  - METS
- Extensible to include many more
  - DIF, NexML, WaterML, CF, NcML, ESML, DDI, MIENS, ...

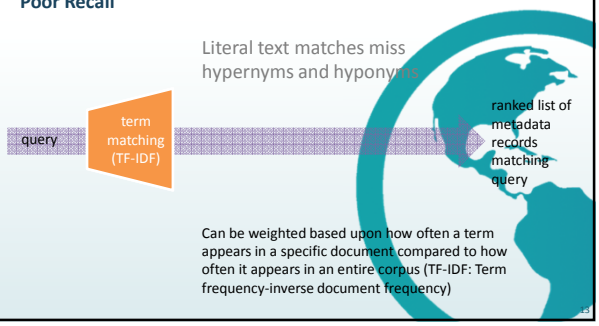
Discover: MN Implications

- Supported science metadata formats
- Accuracy
- Extent
- Data package descriptions

Traditional Term Matching

Poor Precision
Poor Recall

Literal text matches miss hypernyms and hyponyms

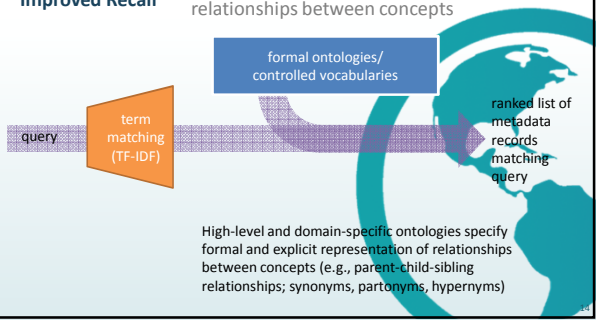


Can be weighted based upon how often a term appears in a specific document compared to how often it appears in an entire corpus (TF-IDF: Term frequency-inverse document frequency)

Adding Formal Ontologies

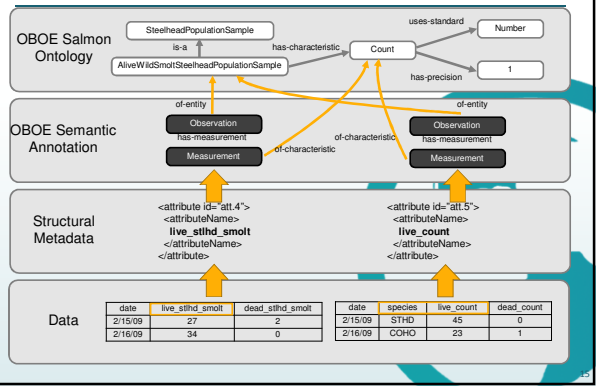
Improved Precision
Improved Recall

Query expansion based on formally-defined relationships between concepts



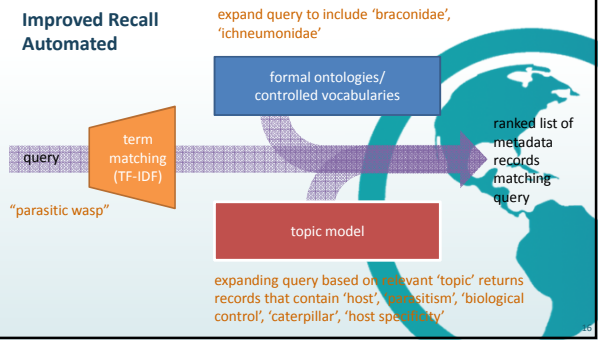
High-level and domain-specific ontologies specify formal and explicit representation of relationships between concepts (e.g., parent-child-sibling relationships; synonyms, partonyms, hypernyms)

Challenge: Linking data to ontology

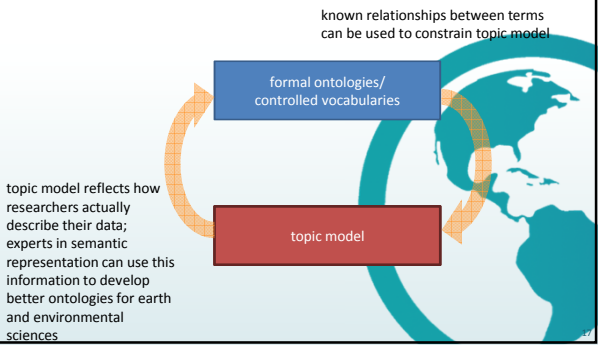


Adding Topic Models

Improved Recall
Automated

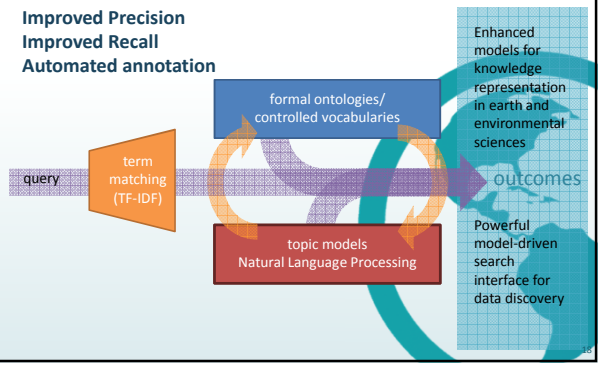


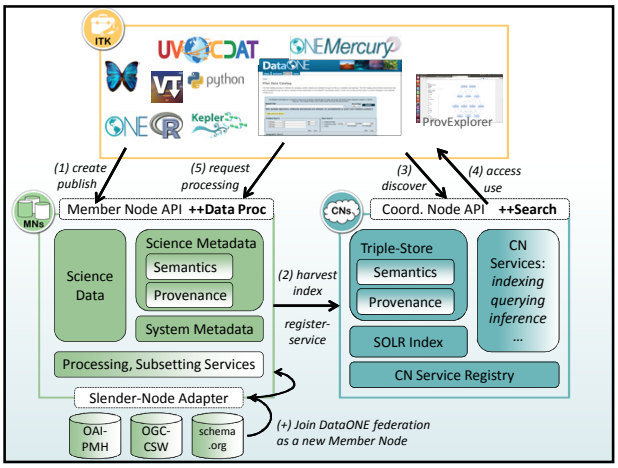
Refinement of Formal Semantic and Topic Models

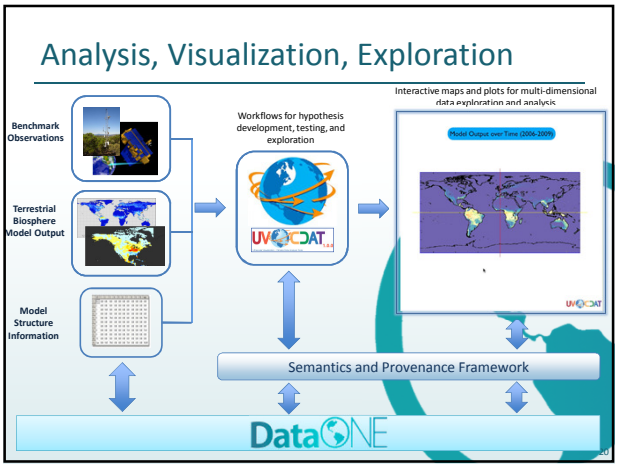


Hybrid Formal Ontologies and NLP

Improved Precision
Improved Recall
Automated annotation




ITK
UVCDAT
ONEMercury
DataONE
python
VT
ONE
Kepler
ProvExplorer
(1) create publish
(5) request processing
(3) discover
(4) access use
MNs
Member Node API ++Data Proc
CNs
Coord. Node API ++Search
Science Data
Science Metadata
Semantics
Provenance
System Metadata
(2) harvest index
register-service
Triple-Store
Semantics
Provenance
SOLR Index
CN Services: indexing querying inference ...
CN Service Registry
Processing, Subsetting Services
Slender-Node Adapter
OAI-PMH
OGC-CSW
schema .org
(+) Join DataONE federation as a new Member Node


Analysis, Visualization, Exploration
Benchmark Observations
Terrestrial Biosphere Model Output
Model Structure Information
Workflows for hypothesis development, testing, and exploration
Interactive maps and plots for multi-dimensional data exploration and analysis
Semantics and Provenance Framework
DataONE