Tools and approaches to facilitate data synthesis

Earth Cube Workshop on “Interoperability of Ocean Time Series Data”

Dr. Mark Schildhauer  (ORCID ID: 0000-0003-0632-7576)
Center Associate, NCEAS, UCSB

Honolulu, HI   Sep 13-15, 2019
**Data Synthesis**

Bringing together *disparate data*, concepts, or theories, integrated in ways that yield new knowledge, insights, or explanations.

National Center for Ecological Analysis and Synthesis
NSF-funded *Synthesis Center*, 1995-2017

Collaborative *Working Groups*
Using ONLY *Existing Data*

*from:*
Data Synthesis Challenges

– **Distributed**: stewarded by many groups, individuals

– **Undocumented**: sparsely and *inconsistently* documented

– **Inaccessible**: varying degrees of presentation and preservation, via Web, paper archives, etc.

– **Heterogeneous**: broad range of topics & measurements (semantics), formats (structure), data access protocols (syntax), data models (theory), etc.
Why preserve and share Data?

Data are the raw materials for science
...re-use: syntheses, AND validating results
...increased scope and robustness of research findings

How to preserve and share Data?

FAIR principles: Findable, Accessible, Interoperable, Re-usable
(Wilkinson et al. https://doi.org/10.1038/sdata.2016.18)
FAIR for DATA SYNTHESIS:

Data that are Findable, Accessible, Interoperable, Re-usable

- **FIND:** Shared, **vocabularies** for vessels, roles, measurements; Globally Unique Identifiers (GUIDs)
- **ACCESS:** Powerful, **open API’s;** Globally Unique Identifiers (GUIDs)
- **INTEROPERABLE:** open source, compatible technologies
- **REUSABLE:** standardized protocols, methods

- **S.O.S.:** **Shared, Open, and Semantic**
  Ontologies to describe terms, schemas
  Machine-assisted discovery, reasoning, integration
Key Technology Needs

...synthesis analyses typically require a broad range of data ranging across multiple scales and encompassing multiple disciplines—heterogeneous in syntax and semantics, distributed across organizations, and often voluminous....

• better **metadata**: structured descriptions of data for discovery/re-use
• **interoperable frameworks** to preserve and access data/metadata
• **better tools** to interact with frameworks through visual and other assistive technologies
• **scientific workflows** to facilitate transparency, reproducibility, sharing, and re-use of analyses
• **robust semantics** controlled vocabularies/ontologies + machine reasoning for finding and re-using data
Approaches

• Data Repository Initiatives (e.g. ERDDAP, DataONE)

• Emerging Standards & Services
  – Schema.org (Google, Yahoo, Microsoft), W3C recommendations (DCAT, RDF/OWL, OWL Time, SOSA, QB4ST), ESIP SWEET, OBO Foundry EnvO, BODC NERC, Darwin Core, EML, Nexus/Newick, FASTA, FGDC, ISO, OGC, CF, etc.
Tools

• Software choices matter:
  – Open source, multi-platform, scalable (HPC, cloud)
  – BETTER: script/code-based software: R, MATLAB, Python, SAS, scientific workflow apps (Galaxy, Kepler, VisTrails, Taverna)
  – NOT AS GOOD: GUI-driven applications: Spreadsheets, black-box compiled applications

• Code repositories needed for preservation of workflows, libraries, functions, etc.
  • GitHub not sufficient for FAIR code?
TOOLS?

Plan  
Collect  
Assure  
Describe  
Preserve  
Integrate  
Discover  
Analyze  

YW  
MATLAB  
DMP Tool  
YW dash  
MATLAB  
YW dash  

Tools:
- YW
- MATLAB
- ONE Share
- ONE Search
- Zotero
- Mendeley
- DMP Tool
- YW dash
Frameworks?

• Large integrative efforts
  – DataONE, Pangaea, EDI, ADC
• These initiatives generally adhere to and promote best-practices, using appropriate community-driven metadata/semantic standards
Solutions for Researchers

• Data Discovery and Access from Multiple Member Nodes

- carbon cycling
- plankton biomass
- ocean nitrogen
- seabird distribution
Research: Traditional Practice

Idea → Design → Deploy/Collect → Organize → Analyze → Write & Publish

Closed, Opaque, Irreproducible, Linear
Research: Open-science, FAIR-enabled

Open, Transparent, Reproducible, Reusable, Dynamic
Lessons for Ocean Time Series Data

• Adopt existing standards where applicable
  • OWL Time, Schema.org, ISO 8601, NERC Vocab

• Interoperate with Existing Frameworks
  • ERDDAP, DataONE API

• Promote FAIR, open source, community-vetted solutions
Lessons for Ocean Time Series Data

• Special querying or sampling needs?
  • Duration, gaps, events, trends: detect, annotate
  • Grain, roll-ups, scaling, synchrony, etc.: compute

• Real-time and static
• Sensor details
Lessons for Ocean Time Series Data

- Special querying or sampling needs?
  - Duration, gaps, events, trends: detect, annotate
  - Grain, roll-ups, scaling, synchrony, etc.: compute
- Real-time and static
- Sensor details

2019-09-13T13:40:51-1000

ISO 8601
Tools Needed for Ocean Time Series Data

- Alignment/mapping tools to “conform” legacy data
- Data entry templates for acquiring new data
- Compatible, open-source analytical tools to process standardized data
- Training in use/interpretation

Owl Time