Data preparation for synthesis - reuse code, collaborate, learn

Environmental Data Initiative (EDI)
LTER ASM, 2018-10-02T13:00-07
Introduction - Challenge - Solution - Vision
Introduction - Challenge - Solution - Vision
Introduction - **Challenge** - Solution - Vision
Introduction - Challenge - Solution - Vision

- Publish
- Plan
- Analyze
- Collect
- Integrate
- Assure
- Discover
- Format
- Archive
- Describe
Introduction - Challenge - Solution - Vision
Introduction - Challenge - Solution - Vision

- Accelerate the efficiency of data processing (IM tasks)
- Reduce redundancy and increase ROI
- Facilitate open and reproducible science
- Elevate IM software as a valued scientific product
Introduction - Challenge - Solution - Vision

- Distributing expertise across a broad community
- Developing software to support best practices through collaboration and community input
- Keeping barrier to participation low
- Serving multiple science domains
- Accommodating a wide range of software maturity
- Containing workflows, packages, single scripts
- Governing committee to facilitate maintenance and development
History - Design - Implementation - Refinement
History - Design - Implementation - Refinement

- **ESIP summer meeting 2017**
  - A hub for sharing and developing code accommodating all levels of software maturity, and requiring minimal oversight.

- **ESIP Science Software Cluster 2017**
  - Implement as lightweight and agile registry rather than repository.

- **RFC 2017**
  - Refine scope
Portal opened 2018
○ Created an OntoSoft portal for the Information Management Code Registry (IMCR).

ESIP cluster 2018
○ Established a space for community discussion, collaboration, and resources.

Hackathon 2018
○ Design and development of software packages to support ingestion of formatted data and QA/QC
History - Design - Implementation - Refinement

- Controlled vocabulary
  - Improve discoverability in w/CV, use extant vocabularies, cross walk with codemeta, test,

- ESIP Summer meeting 2018
  - Highlight software in periodic digests, encourage developer blogging, populate registry

- Harvesting use metrics
  - Working with Ontosoft to track use metrics, currently unsupported

- Registering software
  - On going

- Community building
  - On going
Demo - Wiki - Portal

- Orientation
- News
- Best practices
- Discussions
- General information

bit.ly/IMCRwiki
Demo - Wiki - Portal

- Contribute
- Software ontology (OntoSoft)
- Metadata wizard
- Discovery
- Use
- bit.ly/IMCRportal
Summary - Vision - Participate
Summary - Vision - Participate

- Landscape is changing: large scale synthesis, automation, machine learning
- Increasing the value of science software and receiving attribution
- LTER IM and Science expertise to be leveraged for broad scale ecological DM
- Accelerate the efficiency of data processing (IM tasks)
- Reduce redundancy and increase ROI
- Facilitate open and reproducible science
- Elevate IM software as a valued scientific product
Summary - Vision - Participate

Register your software
- Share your work with others
- Share software you use
- Formalize your software to make it citable and reusable
Summary - Vision - Participate

Controlled vocabulary

- Develop
- Test
- Interoperate
Summary - Vision - Participate

Develop

- Find opportunities to collaborate, improve, effect change
- Work on emerging issues
- Hackathons
  - ingestR
  - qaqcR
  - taxonomyCleanR
Summary - Vision - Participate

- wiki: bit.ly/IMCRwiki
- email: esip-imcoderegistry@lists.esipfed.org
- monthly telecons: 1st Wednesday, 14:00 EDT
- contact: krvander@fiu.edu, colin.smith@wisc.edu
Acknowledgements

NSF grants #1565103 and #1629233
Thank you!
Thank you!
Scope

- Is scope appropriately wide and defined?
  - Information management code:
    - formatting
    - QA/QC
    - metadata
    - archive
    - reuse
    - display
  - Within science domains
Controlled vocabulary

- Categories
  - Function
  - Domain
  - Maturity

- bit.ly/imcrvocab
Development areas

- What are current needs and emerging challenges?
  - Data synthesis within LTER network
  - Serving real-time high frequency data