This is a community notes document

Resources:

IMCR Wiki (Project landing page)
IMCR Portal (Where software is discovered)
IMCR Controlled Vocabulary (Terms related to IM tasks)
LTER Controlled Vocabulary (Terms related to science “disciplines”)
OntoSoft Ontology (Technical information)

General notes:

Does the LTER controlled vocabulary utilize the GCMD Keywords or SWEET ontology? A: No although there is some overlap. Around 2011, when the LTER CV was created only about 20% of commonly used LTER terms were in GCMD. There is some divergence between focuses on Earth Science (GCMD, SWEET) and Ecosystem Science (LTER) that.contributes to the differences.

Activity notes:

Delineating potential:

How else can this resource be utilized?

Would be good to connect tools to the types of data to which it would be applicable - link to data catalog. Ontosoft Ontology does support this capability. But it is a challenging problem due to subtle differences in data (all spreadsheets not the same). Talk with Data Discovery Studio.

Look at alternative ontologies SWO software ontology - LTER list may not be sufficient to the breadth of the community.

What about student contributions of code they have developed? Would they be appropriate? A: yes, for useful code that has some maturity. May be a challenge to find and identify. We are not a repository, just a list, so software would need to be archived elsewhere (e.g., github)

Where are the collaborative opportunities?

NASA EOSDIS is working to catalogue their relevant earthdata tools, and there's an international effort (CEOS) working on doing similar… not sure if linking to those catalogs/repositories would be overwhelming or helpful.

What aspects of our plans need reconsideration?

The scope seems a bit nebulous and overwhelming; perhaps start with a few partner institutions or a narrow community set and progressively build outward from there.

Facilitating development:

What info can be mined from the metadata?

Depends highly on who's metadata; There's so much variability...

What are effective ways of gathering community ideas?

ESIP, AGU, etc. sessions

Tap into/link to/scrape from other agency/community resources (e.g. https://earthdata.nasa.gov/earth-observation-data/tools), though they're likely to be highly specific to the datasets/community for which they were made

How can we support hackathons?

What other ways can we facilitate development?

Test-driving search and discovery:

Do you find what you expect?

Search Relevancy is a big, angry bear that's mauling everybody. If you devise a good, generalizable way solve this, please bring it back to ESIP!

Do the search fields support the content you'd like to search on?

Comments on the vocab structure?

If using a community-agreed controlled vocabulary, can use that as a facet to search or select by.

What terms should be added or removed?

Handling non-generalized code:

How can we support this type of software?

Is there a utility or is it just clutter?

How could it be implemented?

Who vets/quality checks what gets in?

Kristin's Notes:

Bit.ly/imcr-notes
Need audience feedback and comments before going live!
IMCR facilitates discovery and reuse of software for IM use.
Small research groups need IM expertise. Big ones have them already. Elevate IM expertise throughout the environmental and ecological community, because there are increasing expectations to produce openly available and reusable data packages. That are of high quality.
We're targeting small research groups.
Scope: provide software for IM tasks; we emphasize open source software that is community supported. Don't exclude proprietary sources.
Goals: accelerate IM tasks. Simplify discovery and reuse, searchable by task. Return high level information that facilitates fitness for use assessment. Ancillary goal: creating curated

registry, there's a lot of info that can inform future developments. Highlight new opportunities by discovering coverage gaps.
Current status:
Implementation: In Ontosoft: human-friendly interface for discovery. EarthCube.
Curation: ongoing process involving manual discovery and metadata entry. Focusing on R at the beginning, searching CRAN and rOpenSci for packages. Now getting into Python, sciPy. Mathworks in the future. 183 software libraries.
Software by task: assurance, collection, describing, cleaning, quality control – terms used most frequently.
Use this to inform coverage gaps. Identify new development opportunities.
Word cloud of all developers associated with software packages:
Discovery: Ontosoft provides a search interface to discover by core attributes. Author, keywords, implementation language, license, OS, publisher.
Keyword searches are enhanced by keywords organized around the DataOne life cycle.
Domain from section of the LTER CV: Not yet implemented
Software CV: vocab.lternet.edu/
Automated maintenance of software can happen through Ontosoft API. Make sure software are current. Back up the metadata resources. Will have a github repo where all code will live.
Engagement: potential to highlight gaps in coverage.
Hackathons we've done: Very effective! Create a lot more than we could on our own.
Summary: software that runs from local machine or web service; focus on free software; small teams of research groups that don't have form IM expertise and training.

Activities:
1. Delineating potential for this resource
2. Facilitating development: identify gaps in coverage and filling them
3. Test-drive search and discovery in portal: can you find what you want to find? Review IMCR CV? Any additional terms to add? Reorganization
4. Handling non-generalized code: alpha-release code; doesn't below to software package.

Ilya: how to advertise:
Connect it with the types of data it can process. Match software to types of data being published. Many flavors of csv and Excel –
Earth Resource Registry. Data Discovery Studio. Launch jupyter notebook based on the type of data exposed. 1.7 million geoscience datasets. From each one you can launch Jupyter.
SWO software ontology. List of tasks by EDI is controversial. SWO tries to distill tasks at second level. Should we align?

Susanne Remillard: Programs students have written for cleaning their data. She thinks some student scripts would be suitable for IMCR. Cleaning of time-series data. Fairly well developed.
Non-generalized code: Stephan thinks it is of great importance. IMCR is best resource to handle this sort of thing. Also, Ontosoft site is sparse. Need to wrap it better. Needs a little more …. Something before we launch it.
Colin: We started with existing mature packages.
Stephan: What is the threshold for accepting something into IMCR. ESRI2EML code is in IMCR, but nobody can get it to work. Threshold: production ready vs. everything else. Important to scope this project. Stay with libraries, or any code that does any transformation. Other registries of code do exist. NCSA scraped github and saved in database. We could scrape function calls. Scope tightly for libraries we .. Session at ESIP DC meeting about this ….
EarthCube conceptual design project: Geoscience enterprise architecture. Spreadsheet of 700 software packages.
How can we connect to small groups? Gulf of Maine OBFS sites. Corinna: Are there any people who would even be interested in it?
Science Gateways Community or Institute: Catalog of registries.
RDA registry—another place to advertise.
Corinna: What kind of code should be allowed in? At level above stack exchange!
Non-generalized code discussion:
Stephan:
Could Suzanne's student register the software themselves? Yes!
This is not Stack Overflow!
How to organize the content: We could stand up a separate registry for such code: Or add terms to IMCR to indicate production ready or not. Then filter search results accordingly.
Corinna: No to second registry. But don't overwhelm system with crap. It becomes work. We would need moderators.
We would need to find the thousands of pieces of software and figure out how to curate.
Scope: to determine what to harvest from github. Everything with "data management" in title.
Corinna: just link to what Ilya's group has scraped from github. It is in a database somewhere. As time permit, someone could sift through that.
Colin: then use Ontosoft API to put some scraped software into the registry.
Ilya: Google spreadsheet with 700 lines is online somewhere. 2016.
Stephan: People willing to upload have some sense that their software is good. It puts a level on the quality.
Corinna willing to look to see how much clutter comes in once it's advertised. By next year we might need a governance structure.

Ilya: Show utility if from data discovery interface you can launch software from the registry.
Ultimate use case.
Half step toward that goal: neuroscience; operates on certain types of neuroscience tools.
NITRC
Hsieu: A learning opportunity for whoever wants to use it.
Ilya: 470000 notebooks harvested by Bengt.
Susanne:
Corinna: She has a script that turns NTL into ODM. Right for the IMCR? No. Better to collect scripts from all sites in a github and then register that.
Kristin: Put Suzanne's script in github along with input data and sample output.