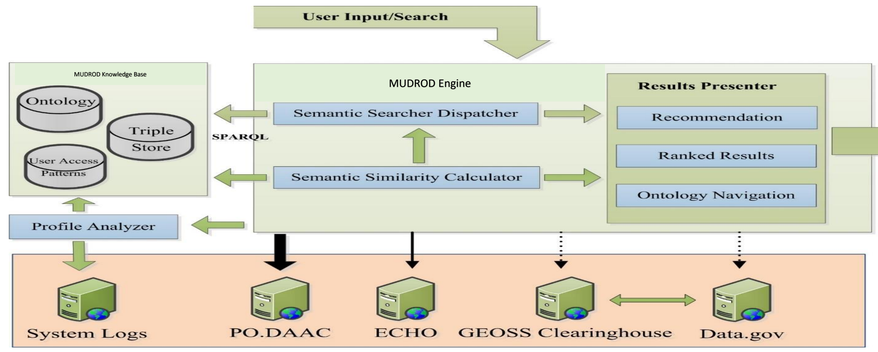
# Mining and Utilizing Dataset Relevancy from Oceanographic Dataset (MUDROD) Metadata, Usage Metrics, and User Feedback to Improve Data Discovery and Access

Chaowei (Phil) Yang, Yongyao Jiang, Yun Li, NSF Spatiotemporal Innovation Center, George Mason University
Edward M Armstrong, Thomas Huang, David Moroni, NASA Jet Proplusion Laboratory

## 1. Introduction

Oceanographic resource discovery is a critical step for developing ocean science applications. With the increasing number of resources available online, many catalogues and portals have been developed to help manage and discover oceanographic resources. However, efficient and accurate resource discovery is still a big challenge because of the lack of dataset relevancy information. We propose a new search engine framework developed by mining and utilizing dataset relevancy from oceanographic dataset metadata, usage and search metrics, and user feedback. The objectives of this project include:

## 2. Architecture& & Objective




I. Analyze web logs to find implicit datasets and keywords relations (current stage)
II. Construct a knowledge base by combining semantics and profile analyzer
III. Improve dataset discovery by 1) better ranked results 2) related dataset recommendations 3) and ontology navigation

## 3. Workflow

**Import PO.DAAC logs**
- Filter out known robots/crawlers
- Remove requests of js, css, img, etc.
- Only HTML requests are imported (timestamp, IP, etc.)
- Input: 4,191,741 requests, output: 297,569 requests

**Import FTP logs**
- Input:3,174,458 ftp output:3,174,458 ftp
- All ftp logs are imported (no user-agent info, and other requests)

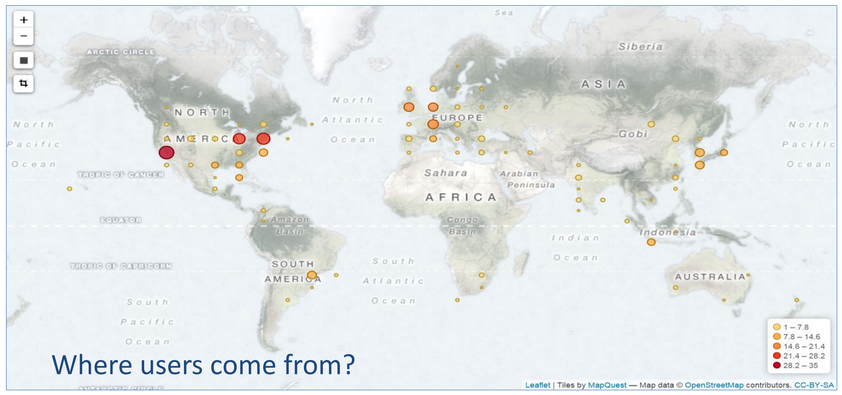
**Sync PO.DAAC web and FTP logs of by IPs**
- Input: PO.DAAC and ftp logs from above two steps
- Filter out IPs that send more than 2 requests/s
- Output: 1) 901,945 logs, 2) unique IPs: 7,536

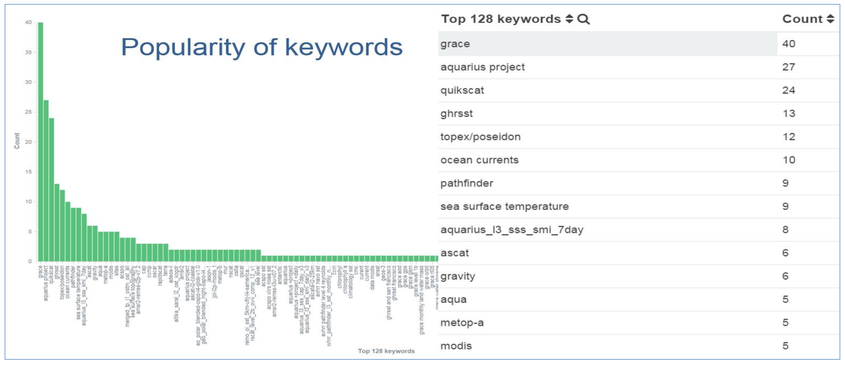
**Connect individual requests into sessions**
- Individual requests are connected based on referrer (Previous page) within the past 10*N mins
- If SessionA – SessionB <10mins, merge them together

**Session filtering**
- Searching request >0
- Viewing request >0
- Downloading request>0

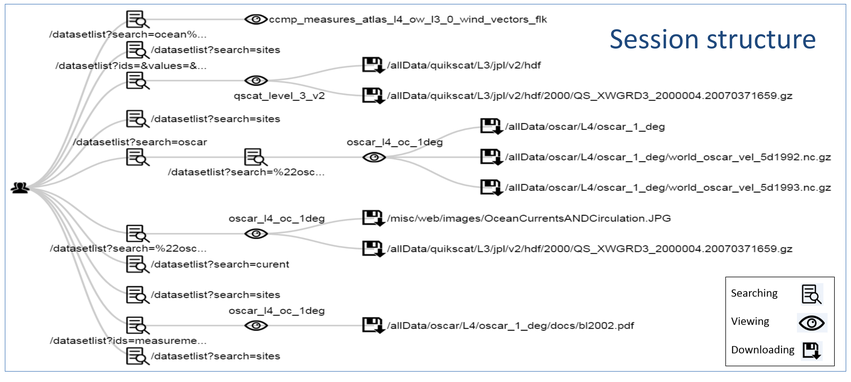
**Session reconstruction**
- Reconstruct session workflow based on referrer
- FTP requests are attached to the nearest viewing request

## 4. Results




| Top 128 keywords | Count |
|---|---|
| grace | 40 |
| aquarius project | 27 |
| quikscat | 24 |
| ghrsst | 13 |
| topex/poseidon | 12 |
| ocean currents | 10 |
| pathfinder | 9 |
| sea surface temperature | 9 |
| aquarius_l3_sss_smi_7day | 8 |
| ascat | 6 |
| gravity | 6 |
| aqua | 5 |
| metop-a | 5 |
| modis | 5 |

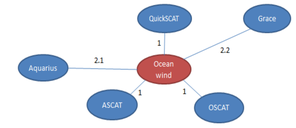



## 5. Next steps

The result will be integrated with semantics.
I. Calculate the probability that keyword A is searched after/before B in a session (Machine learning algorithm-association rules, sequence mining and Markov Chains)
II. Construct OWL using extracted keywords, but the distance is determined by probability
III. Translate searching keywords and calculate similarity
IV. Also, audience could help validate sessions and build future semantics

```
<owl:Class rdf:about="http://localhost/default#QuickSCAT">
   <subClassOf rdf:resource="http://localhost/default#OceanWind"/>
</owl:Class>
```


```
<owl:Class rdf:about="http://localhost/default#dist1">
   <distBetween rdf:resource="http://localhost/default#OceanWind&Grace"/>
   <weight rdf:resource="http://localhost/default#2.2"/>
</owl:Class>
```




## 6. Reference

Liu, K., Yang, C., Li, W., Gui Z., Xia, J., 2013. Using semantic search and knowledge reasoning to improve the discovery of Earth science records: an example with the ESIP Semantic Testbed. International Journal of Applied Geospatial Research.

Aye, Theint Theint. "Web log cleaning for mining of web usage patterns."Computer Research and Development (ICCRD), 2011 3rd International Conference on. Vol. 2. IEEE, 2011.

## 7. Acknowledgement

This project is funded by NASA AIST (NNX15AM85G).