# Search Relevance Progress in NASA

Ed Armstrong and Lewis Mcgibbney
Chairs of the NASA Search Relevance ESDSWG
2016 Summer ESIP Meeting
July 2016

National Aeronautics and
Space Administration

**Jet Propulsion Laboratory**
California Institute of Technology
Pasadena, California

© 2016 California Institute of Technology.
Government sponsorship acknowledged.

# Motivation

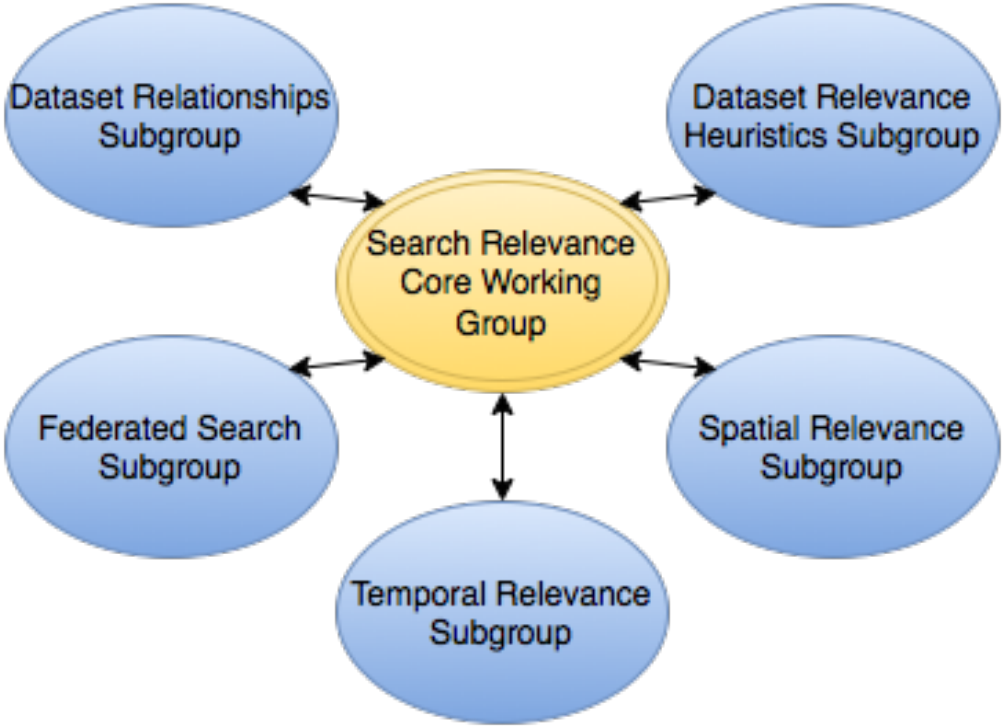
- Search challenges across large and diverse (geophysical type, processing level, format type, quality, tools and services)
  - Thousands of unique datasets, 100 millions of granules
  - Volume and Variety, Velocity and Veracity increasing
  - …..Google has indexed 60 trillion web sites up from 1 trillion in 2008
- Search Relevance 101 (Lynnes and Newman)
  - identifying issues and tests (heuristics) for improving relevance for: usability, spatial and temporal search etc.
- Search served through by the Common Metadata Repository (CMR) and individual NASA earth science data centers

# 2015 Action Plan

## 2015 Working Group Action Plan

| | |
|---|---|
| **Mission Statement**<br>Improve search result relevance for EOSDIS data. | **Stakeholders**<br>• End Users (esp. those who do not know what data are available)<br>• EOSDIS and DAACs by providing a better search experience. |
| **Approach**<br>Subgroups will focus on specific topics:<br>1. Dataset relevance heuristics<br>2. Temporal relevance<br>3. Spatial relevance<br>4. Dataset relationships<br>5. User characterization<br>6. Implementation in distributed search<br>7. Content-based ranking<br>8. Semantic approaches<br>9. Unsupervised learning<br>10. Link-based ranking and relevance<br>11. Granule-level relevance | **Outcomes, Deliverables**<br>1. Recommendation for Essential Metrics (for EOSDIS and DAACs)<br>2. Benchmark dataset of dataset rankings for canonical searches<br>3. Recommendations for Enhancements to Common Metadata Repository<br><br>1 |

# Group Focus Areas



# 3 Core Topic Drivers

- **Investigating essential metrics to improve search result relevance**
  - identify and agree on useful heuristics related to relevance
  - decide on sorting versus scoring
  - identify search-support heuristics to be collected
  - identify performance-criteria heuristics to be collected
- **Investigating enhancements to the CMR to facilitate these searches**
  - investigate how searches can be made within the CMR for e.g. using facet or free text searches
  - investigate the ontology of metadata
  - investigate and differentiate between:
    - time/space resolution
    - time/space coverage
    - time/space repeat coverage
    - an event
    - investigate modeling user history/ behaviors
- **Creating a benchmark dataset of dataset rankings for canonical searches**
  - investigate utilization of linked and semantic data vocabularies within dataset homepages
  - investigate feature co-occurrence present within scientific data sources

# 2015 Subgroup Focus Areas

* Spatial Relevance
  * Improving search relevance based on dataset spatial characteristics
* Temporal Relevance
  * Improving search relevance based on dataset temporal characteristics
* Implementation in Federated Search
  * Improving ranking for results returned from federated (simultaneous search across multiple resources)
* Dataset Relationships
  * Improving search results to return relevant related datasets (cross parameter, cross discipline)
* Dataset Heuristics
  * Establishing metrics and rules of thumb for testing and quantifying search relevance improvements

# External Presenter

| Speaker | Topic |
| --- | --- |
| Dr. Grace Hui Yang,<br>Assistant Professor, Dept. Comp. Sci, Georgetown University | Dynamic Information Retrieval Modeling |
| Dr. Lindsey Spratt,<br>HighFLEET Inc. | HIGHFLEET Semantic Federation Integrated with an Ontologically-Driven Deductive Database System |
| Nathan Clark<br>NASA EED Program Intelligence, Information and Services at the Raytheon Company | Search on Earthdata powered by EOSDIS |
| Dr. Djoerd Hiemstra,<br>Associate Professor, University of Twente | Federated Search |
| Dr. Stefano Nativi,<br>Head of the Division of Florence of the Institute of Atmospheric Pollution Research of the National Research Council of Italy (CNR-IIA) | GEO DAB ranking strategy and applied metrics |

# 2015 Recommendations

* Spatial and Temporal Relevance
  * Spatial and temporal overlap, and intrinsic spatial and temporal resolution are key factors to be used in search result ranking
  * Ranked results from targeted keyword driven regional spatial searches should be improved

* Dataset Heuristics
  * Accuracy heuristics tested using search recall and precision should be implemented.
  * Keyword searches should weighted to favor collections containing controlled science keywords (e.g., GCMD keywords)
  * Search clients should record click through of dataset results

* Implementation in Federated Search
  * Federated searched should utilize normalized discounted cumulative gain (nDCG) to measure topical relevance as the primary metric when merging results from the federated queries

* Dataset Relationships
  * Scientific literature should be mined for data and dataset semantic relations and keyword associations. This information can be used to quantify relationships via ontologies and databases holding triple stores that search engines can eventually leverage.

* Full Recommendations draft document
  * https://docs.google.com/document/d/1XQROrzd4f3HzjHC1vjRwdffkTckJ5raJz5nH79l2bDI

# 2016 ESDSWG Search Relevance II Follow On Work

* Dataset Relationships (existing)
  * Finding and promoting linkages among related datasets. Coupled with Semantic Approaches.
* Dataset Relevance Heuristics (existing)
  * The tests and rules of thumb. All sub group activities feed into this theme.
* Content-based Ranking
* Granule-level Relevance
  * Granule quality, variable content, temporal and spatial relevance are factors
* Link-based Ranking and Relevance
  * Mining user behavior and web site/tool/service interactions.
* Semantic Approaches
  * Using semantics and ontologies to define and quantify domain concepts and intersections among NASA datasets
* Unsupervised Learning
  * Data and keyword mining. Related to Semantic and Dataset Relationships.
* User Characterization
  * User tagging. Related to Granule Relevance.
* Dataset Maturity models
  * Ranking data by maturity

# Summary

* In its first year (2015-2106) the Search Relevancy ESDSWG organized itself into five subgroups that converged on investigating essential metrics to improve search result relevance and identifying enhancements to the CMR to facilitate improved search.
* The results of the Year 1 work are presented in the **ESDSWG Search Relevance Recommendations** which is being made available as the ESDSWG community as the **WG Editor's Draft April 2016**
* Search Relevance II will continue in 2016-2107 to review additional topic areas including a subset of (TBD):
  * Granule Relevance, Link-based Ranking, Semantic Approaches, Unsupervised Learning, User Characterization, Dataset Maturity Models and User Intent Modeling
* 2016 Fall AGU Session IN030**. New Approaches to Data Discovery Across Geoscience Domains**

# Group Contributors

* James Norton
* Kim Whitehall
* Beth Huffer
* Vicky Wolf
* Chris Lynnes (ESDIS POC)
* Hampapuram Ramapriyan
* Siri Jodha Khalsa
* Kyle Huffman
* Petr Votava
* Alicia Aleman
* Eli Pauli
* Daniel Da Silva (chair, resigned)
* Ed Armstrong (chair)
* Lewis Mcgibbney (chair)

# Resources

* [0] http://go.nasa.gov/21SwufF
* [1] http://gcmd.nasa.gov/
* [2] https://earthdata.nasa.gov/about
* [3] http://go.nasa.gov/25H2XHr
* [4] http://go.nasa.gov/1RUA4Az
* [5] http://bit.ly/25H4BsO
* [6] http://go.nasa.gov/1PQDSy5

# Acknowledgments


* The members of the Search Relevancy WG acknowledge their individual funding sources for supporting their participation in this working group

* Please direct any correspondence, feedback or WG participation to the WG mailing list at esdswg-search@lists.nasa.gov.