# Search Relevance Recommendations for Earth Science

## Status of this Memo

This memo provides information to the NASA Earth Science Data Systems (ESDS) community. This memo does not specify an ESDS standard of any kind. Distribution of this memo is unlimited.

## Change Explanation

This document is not a revision to an earlier version.

## Copyright Notice



## Abstract

This document contains a series of recommendations developed by the NASA Earth Science Data System Working Group (ESDSWG) on Search Relevance and User Characterization over a three-year period from 2015-2018 and is intended to improve search relevance ranking from NASA search engines and related interfaces. The primary stakeholders of this document are considered to be DAAC managers, Evolution and Development (EED) contractors and other such persons who are responsible for deciding on the interfaces, tools and services needed by end users to discover NASA's vast relevant Earth science datasets. Additional stakeholders for this document include

- data producers and data publishers (from the scientists end users' category that includes instrument teams and scientists who are creating custom data sets for their research and DAACs),
- data consumers (scientists and modeler end users' categories that include researchers directly studying and analyzing observations, as well as climate modelers and weather forecasters),
- software developers who are responsible for discovery and access tools (that include EED contractors and data product and service developers at the DAACs)

**Table of Contents**

## 1 Introduction

The ESDSWG Search Relevance working group (WG) was initiated after the 2015 ESDSWG annual meeting. Primary WG motivations related to inconsistencies of returned search results from various NASA resources such as Earth Observing System (EOS) Clearing House (ECHO), Global Change Master Directory (GCMD), and others, hence the WG tasked itself to review, strategize, and develop recommendations for improving search relevance rankings from NASA search engines such as the Earthdata Search (ES) of the Common Metadata Repository (CMR), and related interfaces.

Readers of this document will benefit from a concise set of search relevance recommendations drawn from several areas of study including spatial and temporal relevance, dataset relevance heuristics, semantic dataset relationships, federated search, content-based optimization for commercial search engines, and user characterization. These recommendations can be used to dramatically improve information retrieval software systems such as search engines.

The remainder of this document lays out the 14 recommendations made by the WG during the 2015-2018 period of activity. Readers should note that as of early 2019, some recommendations have motivated extended work efforts outside of ESDSWG. For example, our recommendations related to content-based optimization for commercial search engines are being further developed through ESIP's Semantic Technologies Committee.

## 2 Recommendations

Table 1 below groups recommendations by topic for improved navigation

| Topic | Recommendations(s) |
|---|---|
| spatial and temporal relevance | 1 - 6 |
| dataset relevance heuristics | 7 - 9 |
| semantic dataset relationships | 10 |
| federated search | 11 |
| content-based optimization for commercial search engines | 12, 13 |
| user characterization | 14 |

Note, the WG did not propose any recommendation on the topic area of granule relevance but decided to leave it open to future ESDIS activities.

The structure of recommendations is motivated by modern, well understood W3C practices. This follows the *challenge, description, intended outcome, possible approaches to implementation* and finally *how to test* logic. This has been determined to be easy to interpret, clear and actionable. The reader should however be aware that the *how to test* sections are not provided for each recommendation, in which cases the testing criteria should be interpreted from the *intended outcome* and *approach to implementation*.

## 2.1 Recommendation 1: Spatial overlap heuristic

The Challenge

Although currently in the CMR, end users are allowed to search on keywords and to enter spatial constraints, the relevance rankings for results are based only on the keyword match. The challenge is to improve returned results by ranking higher those datasets (or granules) that overlap a specific region in space and/or have a significant percentage of usable data in the specific region, over those datasets that only match keywords.

The Recommendation description

We recommend to calculate the overlap between a user-supplied coordinate bounding box and the datasets returned that match the keyword search related to the physical parameter the end user desires.

Intended Outcome

Rank the datasets based on overlap percentage to a spatial query. This overlap can be calculated by the physical bounds of the intersection of the data to the region of interest. It can also be calculated as the percentage of usable data contained in the regions or spatial features when performing granule searches.

In considering an implementation as suggested below, it is however important for a search engine provider to consider and leverage the wide availability of standardized spatial predicates which can be used to associate and rank results. The OGC's *Implementation Standard for Geographic Information – Simple feature access – Part 1: Common architecture* (OGC, 2011), specifically Subclause *6.1.2.3 Methods for testing spatial relations between geometric objects* identifies the following geometry subtypes

- **Equals**; if a geometric object is "spatially equal" to another geometry.
- **Disjoint**; if this geometric object is "spatially disjoint" from another geometry.
- **Intersects**; if this geometric object "spatially intersects" another geometry.
- **Touches**; if this geometric object "spatially touches" another geometry.
- **Crosses;** if this geometric object "spatially crosses" another geometry.

- **Within;** if this geometric object is "spatially within" another geometry.
- **Contains**; if this geometric object "spatially contains" another geometry.
- **Overlaps**; if this geometric object "spatially overlaps" another geometry.
- **Relate**; if this geometric object is spatially related to another geometry by testing for intersections between the interior, boundary and exterior of the two geometric objects as specified by the values in the intersection pattern matrix. This returns false if all the tested intersections are empty except exterior (this) intersect exterior (another).
- **Locate Along**, returns a derived geometry collection value that matches the specified m coordinate value. See Subclause *6.1.2.6 "Measures on Geometry"* for more details.
- **Locate Between**; returns a derived geometry collection value that matches the specified range of m coordinate values inclusively. See Subclause *6.1.2.6 "Measures on Geometry"* for more details.

To aid in the approach below, the methods defined above as taken from *Part 1: Common architecture* are implemented in SQL in the *Implementation Specification for Geographic information - Simple feature access - Part 2: SQL option* (OGC, 2010). Specifically, see Subclause *7.2.8 SQL routines on type Geometry.*

Possible Approach to Implementation

The CMR Search Engine can be configured to determine the degree of spatial overlap in real-time searches. However, complex polygons defining the region of interest will affect the speed of these computations.

## 2.2 Recommendation 2: Sorting by regional area

The Challenge

Often a search will target a specific region via a free-text keyword search (e.g., Amazon River, West Africa). Those targeted matches should be ranked highly.

The Recommendation description

This recommendation is related to the spatial overlap heuristic but is driven by keyword match vs. spatial overlap match. We recommend an update to the database schema to facilitate indexing datasets with regional area tags.

Intended Outcome

Improved ranking and searching by regions of interest via a free-text keyword search.

Possible Approach to Implementation

Datasets can be tagged to make an explicit association with specific spatial regions e.g., Atlantic Ocean, Amazon River, etc. Consideration should be given for how a keyword representing a regional tagged location will work together with a spatial overlap parameter and ultimately which one, if any will be given more weight in the overall query. This is a detail we leave to the search engine implementation as undoubtedly the weighting may change over time due to the variability in the definition of regional areas. An example here would be a glacier receding or a flooding plane expanding. In these cases, two or more search criteria *may* conflict so this is something the search engine implementors should be aware of.

Finally, due to the variability issue raised in the previous paragraph user or machine-provided tags should always be treated with caution. Factually incorrect or inaccurate tags can negatively affect search relevance on a number of levels. This can be the result of tagged information being incorrect to begin with or simply expiring in factual accuracy over time as suggested in the previous example.

### 2.3 Recommendation 3: Spatial resolution heuristic

The Challenge

The spatial resolution for datasets is important to scientist end users as the native pixel size or resolution of the dataset (gridded and station data) determines the type of Earth science phenomenon which can be resolved.

The Recommendation description

We recommend an update of the database schema to facilitate indexing datasets with spatial resolution information; this information should include units.

Intended Outcome

Datasets are ranked and sorted via their intrinsic spatial resolution. Similar as to the intended outcome guidance provided in **Recommendation 1**, we also strongly encourage the user to consider the full spectrum of spatial resolutions when working towards an implementation of this recommendation. Examples would include resolution close to, finer than, and coarser than the resolution wanted by users. The WG did not investigate this topic in more detail but recognizes the importance.

Possible Approach to Implementation

Datasets must be indexed or tagged for their resolution specifications within the database. This will allow faceted search. The ability to parse queries and perform unit conversion e.g. coordinate

reference system mappings such as from 1km resolution to 0.01-degree, may also be necessary depending on varying unit notation and user competency.

## 2.4 Recommendation 4: Spatial and temporal proximity to an "event" heuristic

The Challenge

Searches for data, especially in the public and education categories, often take place in the context of an event such as an earthquake, volcanic eruption, or a hurricane. It should also be noted that event geometries come in various forms e.g. geographic point, multi-point, polygon, multi-polygon, etc. so it may be the case that a query matches a grouping of event geometries such as a hurricane track over time. In this scenario, the end user is interested in acquiring data from multiple sources, including various satellite sensors, with spatial and/or temporal proximity to the event. This use case has both real-time and historical applications.

The Recommendation description

Create virtual collections for events to facilitate these types of keyword searches.

Intended Outcome

Those datasets in close spatial and temporal proximity are ranked with higher relevance. For commentary on standard geographic information relating to simple feature access we strongly advise the reader to consult (OGC, 2011) then (OGC, 2010).

Possible Approach to Implementation

A virtual collection that potentially spans data products across multiple data centers could be created and searched by event name, location, or date.

## 2.5 Recommendation 5: Temporal overlap heuristic

See ***Recommendation 1: Spatial overlap heuristic*** for general consideration of the time domain. We do however strongly recommend the reader consult (Allen, 1983) which defines standardized temporal linear algebra comprising thirteen basic relations between time intervals that are distinct, exhaustive, and qualitative. These 13 relations and the operations on them form *Allen's interval algebra* which is accepted to be the defacto mechanism for calculating 1 dimensional time intervals and hence overlap in this context.

### 2.6 Recommendation 6: Temporal resolution heuristic

See ***Recommendation 3: Spatial resolution heuristic*** for consideration of temporal rather than spatial resolution.

### 2.7 Recommendation 7: Performance metrics suite

The Challenge

Without metrics, it is impossible to know whether or not applied heuristics, recommended within this document or otherwise, are improving search relevance.

The Recommendation description

We recommend to implement a suite of performance metrics across all data providers within EOSDIS. The availability of these performance metrics is critical for undertaking comparisons with the aim of establishing baseline relevance rankings.

Intended Outcome

A test set and scoring metrics that can be used to establish baseline performance as well as assessing change in overall relevance when implementing heuristics.

Possible Approach to Implementation

*Precision* and *recall* are very well established Information Retrieval (IR) metrics which establish the performance of a particular computation. The metrics can be calculated as follows:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

**precision** is defined as the number of correct results divided by the total number of returned results and **recall** is defined as the number of correct results divided by the number of results that should have been returned.

How to Test

1) Establish test cases. For example: keyword search = ***ozone***
   a) Ask data providers for common keyword searches. Note: the following data providers may be skewed towards users consuming large scale data as opposed to small scale in situ data.
      i) GES DISC
      ii) LPDAAC
      iii) NSIDC
      iv) ASDC
      v) FIRMS
   b) Based on the availability of keywords, conduct CMR searches via
      i) Free-text search
      ii) specific search; experiments have shown that the '***ozone***' test (as described above) had a provider constraint at the free-text level which results in most responses being provider-centric
      iii) temporally constrained search; may prove a useful additional query parameter if results for particular queries contain old datasets for example
   c) Get 'top dataset' input from each data provider and compare using the metrics given above
   d) Work with a group with expertise in IR search assessment

## 2.8 Recommendation 8: Science Keyword Search Heuristic

The Challenge

Leverage the distinct vocabulary associated with Earth science data to rank returned search results; aim to favor results matching the science keywords as this will increase relevance by avoiding matches that are not directly related to the subject of the search query.

The Recommendation description

Science keywords represent a controlled vocabulary that captures the essence of the data in a collection. Such a resource can be utilized as a heuristic to rank returned results. More specifically, datasets ranking from keyword free-text searches can be weighted to favor datasets returned matching science keywords over those that contain the search keyword anywhere else in the metadata record.

Intended Outcome

In order to satisfy this heuristic, the search keyword(s) would have to describe the actual measurement/physical parameter(s) in the data, such as '***ozone***'. Collections that match the keyword(s) elsewhere, e.g. in the instrument name, would not satisfy this heuristic. Collections with search keywords in the description would rank lower than collections with search keywords in controlled vocabularies(e.g. the Science keywords).

Possible Approaches to Implementation

Currently CMR keyword searches are implemented using an Elasticsearch (Elasticsearch, 2019) *function score query*. This query consists of a set of weighted filters that are applied to all collections that match the primary query. The filters check for matches of search keywords against specific fields, and the CMR weights can found at the overloaded function `def default-boosts` in https://bit.ly/2I7xEm4 . The weights of all the filters that pass for a given collection are multiplied together to produce a relevance score.

$$relevance = \prod_{i=1}^{n} w_i F_i$$

where $w_i$ is the weight of the *ith* filter and

$F_i =$ 1 if the *ith* filter passes, 0 otherwise.

Because the weights are all greater than 1, passing filters always *increase* the overall relevance score.

Function score queries can perform other functions beyond simple weighting, such as decay functions or mathematical operations on numeric fields. This approach could be used for simpler heuristics, such as a heuristic based on collection recency, in which the contribution to the relevance would be a simple function of the collection end date.

The drawback to this approach is that it is limited to the functionality provided by function score queries, which means filters on field values and scripting basic mathematical functions on numeric fields. Also, any fields used in scripting would need to be stored, increasing the size of the index.

An alternative approach is to incorporate relevance heuristics into an Elasticsearch plugin similar to that currently used by the CMR to compute spatial filtering, as shown in Figure 1.

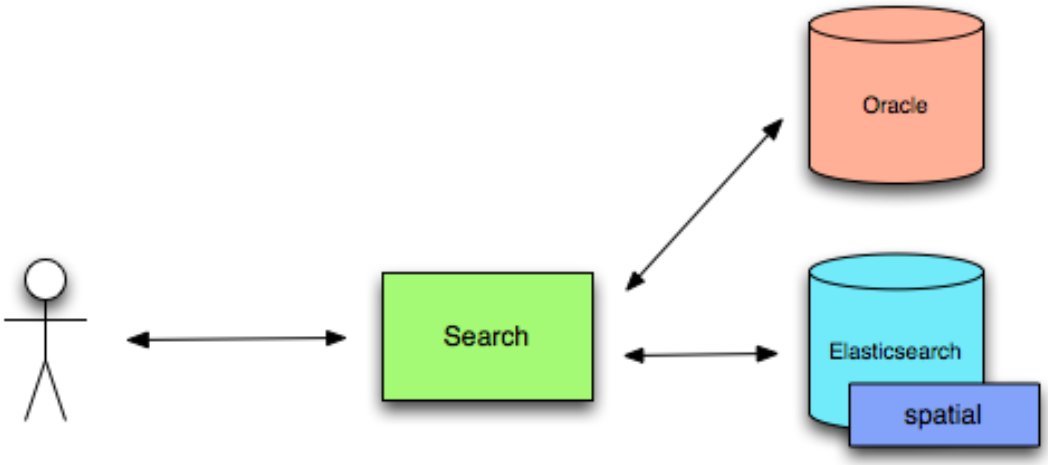


**Figure 1** – Example of a client engaging in CMR Spatial Filtering via a Custom Elasticsearch Plugin

The spatial plugin filters out search results that are outside of the search area specified in the search query by analyzing all the search results matching the rest of the query. So a search for '***ozone***' with a spatial constraint would only match those results related to '***ozone***' *and* in the specified area.

In a similar way, a custom relevance plugin could be added to Elasticsearch to incorporate relevance heuristics as illustrated in Figure 2.

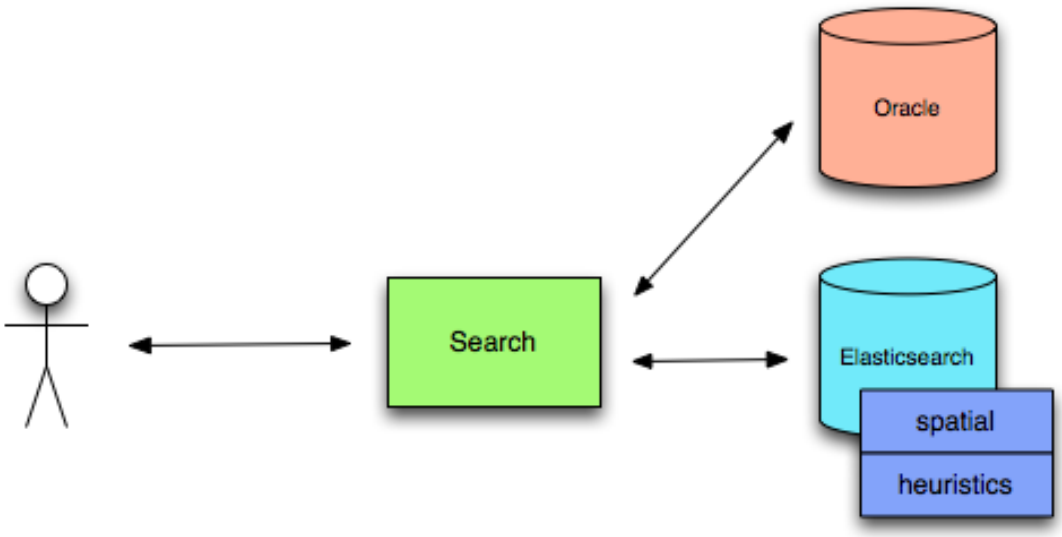


**Figure 2** - Utilizing Custom Plugins which Implement Heuristics to Compute Search Result Relevance.

Like this spatial plugin, this plugin would operate on every search result and compute the overall relevance score for each document. This approach is more flexible than the current function score query approach as it allows *any* function to be used to implement any desired heuristics.

How to Test

Metrics can be computed using the test set constructed in ***Recommendation 7: Performance metrics suite*** using the new weighting; these metrics can be compared to the baseline metrics established in that recommendation.

### 2.9 Recommendation 9: Collect end user behavior data using the Earthdata Search Client

The Challenge

Collecting end user behavior data will enable system and algorithm improvements that will enhance search precision and recall. This offers a method for the DAACs to facilitate discovery especially of interdisciplinary datasets. These metrics would be useful for multiple purposes and provide a check on the assumptions of those developing the search information and aggregation with data on the actual practices of the user community.

The Recommendation description

We recommend collecting end user behavior starting with the foundational results from the ES to record the end user's click events after exercising the returned search results.

Intended Outcome

To provide recommendations of datasets to end users, in a similar manner that eCommerce platforms such as eBay, Amazon, etc. offer supplementary information to end users that selected *X*, and then also selected *Y and Z*.

Possible Approach to Implementation

Combine data from the CMR/ES metrics with the User Registration System (URS) metrics, possibly utilizing ESDIS Metrics System (EMS), to analyze the datasets that are downloaded by the same users, particularly within a short period of time such as a day. The retrieved metrics can be compared with results related to dataset relatedness found through scientific literature mining to determine datasets used together in the published abstracts/papers. See ***Recommendation 10: Identify relatedness between datasets through text mining of scientific literature*** for more details on scientific literature mining.

## 2.10 Recommendation 10: Identify relatedness between datasets through text mining of science literature

The Challenge

The challenge for Earth science is to obtain information that can be used to determine if two or more datasets are relevantly related and to do so in a manner that will scale. This is made more challenging by the fact, noted above, that strictly statistical analyses may not likely to yield useful results, given the relatively small user base, and that the criteria for relatedness between datasets in Earth sciences are dependent on the purpose of the search. For example, given the physical relationship between aerosol optical depth and air pressure, a dataset containing air pressure measurements and a dataset containing Aerosol Optical Depth (AOD) measurements, may be more relevantly related than two datasets both containing AOD. In the former case, variable dependency defines the relatedness between the datasets, while in the latter case, variable similarity defines it. Alternatively, a data scientist conducting multivariate analyses to detect non-obvious relationships among many Earth system phenomena variables including AOD will most likely be interested in finding datasets that lend themselves well to comparisons based on the structure of the datasets, (e.g. spatial and temporal resolution, spatial and temporal coverage - i.e., parameters that are specific to the dataset, or the instrument, or the satellite, as opposed to the physical relationships between science parameters). All of these would seem to constitute reasonable criteria for deeming two or more datasets relevantly related, but only knowing the context of the particular user's search session can determine which criterion should prevail.

While the challenge that context-dependency presents is significant, the group converged on the view that information about relevant relationships between datasets may be derivable from research articles and other textual documentation, which can not only provide data about related datasets, but can also provide useful context for understanding the nature of the relationships. Information embedded in scientific literature, for instance, can relay how scientists are in fact using datasets, which datasets they are combining in their research, and for what purpose they are doing so. Information embedded in Data Product Guides and ATBD's can relay how higher-level datasets were collected and/or processed, the provenance of the data, and the structure of the data model.

The Recommendation Description

We recommend that these types of documents be mined to identify salient relationships and express those relationships as semantic annotations on datasets, providing machine-readable information about variable dependencies, spatial and temporal structures, usage, etc. that can be accessed by search services to provide recommendations to users about other datasets that might be of interest.

Intended Outcomes

The purpose of applying Natural Language Processing (NLP) techniques to textual documents is to 1) extract information about existing physical and technical relationships between datasets being used in scientific research, and 2) generate semantic annotations expressing those relations, which will become part of the metadata records for the datasets referred to in such texts. Search services will be able to leverage these annotations to recommend relevantly related datasets to users who express interest in particular datasets, much like Amazon provides recommendations to their users. NLP-derived semantic annotations will improve the links between datasets and keywords that map to data usage concepts, variables dependencies, and other relevance criteria.

Possible Approaches to Implementation

Mine external academic literature (for example) to identify relationships between datasets and the types of research and science disciplines they are used for. Datasets that are consistently used for specific disciplines can be tagged as highly relevant to that discipline.

Mine academic literature to identify relationships between science parameters that are consistently used together (e.g., air temperature + surface precipitation) and tag data sets with related parameters to help support recommendation services. Utilize linked and semantic data vocabularies within dataset homepages to identify various salient properties of datasets extracted from literature. For example, if a dataset is consistently used for a certain type of research purpose or science discipline, the dataset's landing page could include the purpose or discipline as a schema.org keyword

This mining of scientific literature approach can be used to identify specific research applications for Earth science datasets and provide machine readable summaries to help users identify datasets relevant to their own research interests. Researching tools and semantic technologies in used in the health sciences, in particular those which address data management problems that are generally relevant to scientific research, should be pursued in the coming year. Semantic MedLine (2012), for example, developed at the National Library of Medicine, parses research articles and extracts "semantic predications" (subject-predicate-object triples) from text. Used in conjunction with the Unified Medical Language System (UMLS) Metathesaurus and the UMLS Semantic Network, it can meaningfully identify and classify the arguments in the extracted triples by mapping noun phrases to Metathesaurus concepts and verb phrases to Semantic Network predicates. In the natural sciences, Semantic Web for Earth and Environmental Terminology (SWEET, 2018) is a set of ontologies constructed from the set of keywords in the GCMD that can be leveraged in semantic NLP. SWEET is an open source middle-level ontology that allows users to add a domain-specific ontology using the components defined within. There are numerous other promising open source NLP tools that should also be investigated, including (Stanford's CoreNLP, 2018), (Apache OpenNLP, 2018 ) and (Natural Language Toolkit, 2018).

## 2.11 Recommendation 11: Utilize Normalized Discounted Cumulative Gain as the Primary Measure of Topical Relevance

The Challenge

Not every search scenario is currently facilitated by a system which provides results from one source. Examples of this reasoning include, but are not limited to, data location and logistical-/ performance-related issues with data movement, data intellectual property rights, data access controls, security and authorization restrictions, etc. It is therefore entirely likely that search may be conducted in a federated manner where a query broker acts as an intermediate in order to provide a federated response to any given query. With a specific focus on the issue of *results merging*, once a merged, singly ranked list has been returned to the user of a particular query, how do we measure the topical relevance of results which have come from different underlying data sources?

The Recommendation description

Utilize *Normalized Discounted Cumulative Gain* (nDCG) (Jarvelin, K. and Kekalainen 2000, 2002) as the primary measure of topical relevance within a federated search system. nDCG measures the performance of a recommendation system (in this case a federated search system) based on the graded relevance of the recommended entities.

Intended Outcome

Primary outcome is to derive a score of between 0.0 - 1.0 (with 0.0 being low and 1.0 being high) representing the ideal ranking of entities presented within the merged singly ranked list.

To evaluate how the behavior of federated search system components affect the merged, singly ranked list which is returned to a user. nDCG permits comparative approaches to execution of federated queries, collection selection, collection representation and results merging.

Possible Approach to Implementation

The approach involves calculating a DCG for each data source (otherwise known as a *vertical*). The requirement to calculate the graded relevance of results returned from each vertical which participates in the federated search, necessitates the parameter ***k*** which denotes the maximum number of entities that can be recommended. The premise of DCG is that highly relevant documents appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result.

Based on this premise, the DCG implementation below places stronger emphasis on retrieving relevant documents as this formula is commonly used in industry, including major web search companies:

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2 (i + 1)}$$

Where

***k*** denotes the maximum number of entities that can be recommended

As search result lists vary in length depending on the query, comparing a search engine's performance from one query to the next cannot be consistently achieved using DCG alone, so the cumulative gain at each position for a chosen value of ***k*** should be normalized across queries. This is done by sorting documents of a result list by relevance, producing the maximum possible DCG till position ***k***, also called Ideal DCG (IDCG) till that position. $IDCG_k$ is the maximum possible (Ideal) DCG for a given set of queries, documents and relevance's. For a given query, nDCG is then computed as follows

$$nDCG_k = \frac{DCG_k}{IDCG_k}$$

The nDCG values for all queries can be averaged to obtain a measure of the average performance of a federated search engine ranking algorithm. It should be noted that in a perfect ranking scenario, the $DCG_k$ will be the same $IDCG_k$ producing a nDCG of 1.0.

It should be noted that the most significant barrier to a meaningful implementation of nDCG is the unavailability/inability to obtain domain experts who can adequately generate graded relevance assessments for returned documents within selected verticals. Such an exercise is not however unique to the Federated Search agenda, the lack of known expertise in this area across ESDIS generally is a known issue.

## 2.12 Recommendation 12: Dataset Landing Pages (DLP) should be improved with structured data markup to support discoverability by commercial search engines

The Challenge

It is widely recognized that many users, when looking to retrieve NASA science data, first consult their favorite commercial search engine rather than going straight to the source e.g. the DAAC which hosts the data. With this in mind, it is clear that NASA needs to connect more with how DLP and subsequently dataset discovery, is improved within commercially motivated search engine rankings.

The Recommendation Description

We recommend that the CMR, DAACs and any other public facing infrastructure serving NASA data define a strategy for improving the Structured Data Markup (SDM) of, in particular, DLP for the purpose of improving dataset discoverability and commercial search engine rankings.

Intended Outcome

Each NASA DLP should have an accompanying Rich Text Snippet (RTS) which provides search engine users with a very precise, up-to-date, streamlined overview of some of the dataset characteristics. This differentiates DLP from other search engine results. subsequently reducing confusion about the authoritative source for the correct information. Additionally, this will enable enhanced integration with commercial search engine SDM standards and practices for publishing (geospatial) data on the Web.

Possible Approaches to Implementation

DLP should be improved with SDM such as the schema.org Dataset (2012, 2018) and DataCatalog (2018) types to support discoverability by commercial search engines such as Google, Bing and Yahoo!. The schema.org vocabulary, one of many SDM options, is an open community effort to promote standard structured data in a variety of online applications. This topic is particularly

relevant due to the huge number of datasets that have been made public in recent years, with the resulting uptake in dataset discoverability. *Structured data* refers to kinds of data with a high level of organization, such as information in a relational database. When information is highly structured and predictable, search engines can more easily organize and display it in creative ways, making it more easily visible and understandable to prospective consumers. SDM is a text-based organization of data that is included in a file (e.g. any given DAAC dataset landing page or dataset record served through CMR) and served from the web.

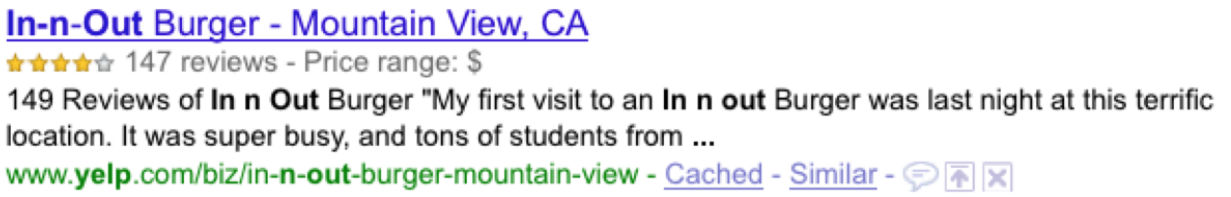


Figure 3: Result snippet with rating, price range, and review

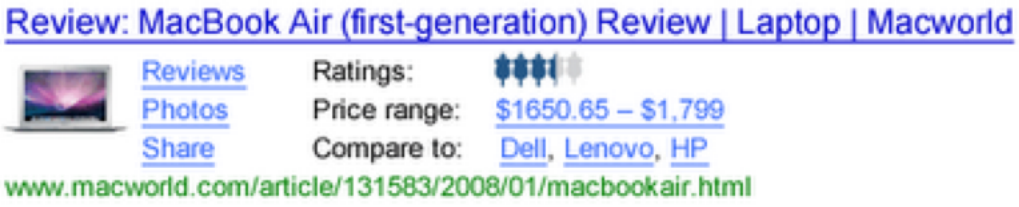


Figure 4: Result snippet with formatting, links, image, and comparative information

Services and software such as Google's *structured data testing tool* (Google, 2018a), Webmasters Search Console (Google, 2018b), and Apache Any23 (Apache Any23, 2018) provide mechanisms for checking how much and to what extent documents served over the web contain SDM. These are excellent resources which can be used to assess how well search engines are able to interpret SDM associated with NASA datasets with the aim of providing custom search functionality for individual collections, datasets and granules. Custom search functionality such as RTS https://github.com/ESIPFed/science-on-schema.org has become possibly the most appealing mechanism for promoting search results within search engine rankings. RTS helps users find the right page by showing them a snippet - a small sample of content that gives search users an idea of what's in the web page. Figures 3 and 4 above show examples of how rich text snippets enhance and advertise specific search engine results.

As of this writing (2019), it should be noted that due to renewed, collaborative efforts between the Search Relevance Working Group and Google, Inc., **all CMR landing pages for collections currently contain SDM which is being harvested by Google**.

Examples of schema.org markup including Microdata include the following

- *itemscope*, *itemtype*; the former stating that the markup itself is semantically annotated (and that the SDM should therefore be interpreted accordingly) and the latter describing the item and its properties context. The following example is a snippet of XHTML stating that the values contained within the unordered list are part of a Dataset, e.g.,

```
<div class="row content collection" itemscope itemtype="http://schema.org/Dataset">
```

- *itemprop*; actual properties associated with a dataset. The following XHTML snippet displays an associated dataset ***name, alternateName*** and ***version***, e.g.,

```
<meta itemprop="name" content ="PODAAC-AQR50-3T7CS"/>
<meta itemprop="alternateName" content ="AQUARIUS_L3_SSS_CAP_7DAY_V5_1"/>
<meta itemprop="version" content="1"/>
```

- *datetime*; dates and times can be difficult for machines to understand. Consider the date "04/01/11". Does it mean January 04, 2011? Or April 1, 2011? To make dates unambiguous, use the time tag along with the datetime attribute. The value of the datetime attribute is the date specified using ISO 8601 compliant date times. The HTML code below specifies the date range unambiguously from August 25, 2011 to June 8, 2015 in ISO 8601 format.

```
<time itemprop="temporalCoverage" datetime="2011-08-25T00:00:00.000Z/2015-06-08T00:00:00.000Z">
```

- Content; sometimes, a web page has information that would be valuable to mark up, but the information can't be marked up because of the way it appears on the page. The information may be conveyed in an image (for example, an image used to communicate spatial or temporal coverage, sensor characteristics, etc.) or it may be implied but not stated explicitly on the page (for example, the spectral operating range of a particular instrument). In these cases, the *meta* tag can be used along with the content attribute to specify the information. Consider this example; the following SDM shows a spatial coverage of a global dataset from -90.0, -180.0, 90.0, 180.0 .

```
<meta itemprop="spatialCoverage">
  <div vocab="http://schema.org/" typeof="Place">
    <div property="geo" typeof="GeoShape">
      <meta property="box" content="-90.0 -180.0 90.0 180.0" />
    </div>
  </div>
</meta>
```

The above provides just a sample of the possibilities for utilizing SDM to improve interpretation and ultimately relevance ranking of NASA datasets within commercial search engine results. Both temporal and spatial overlap have been identified as primary targets for implementation of relevance ranking in the ES. We strongly recommend that DAACs explore SDM for spatial and temporal coverage at a minimum, and that ESDIS explore the possibility of making use of such mark-up as well.

The Content-based Optimization for Commercial Search Engines Subgroup also experimented and prototyped use of CoverageJSON (CoverageJSON, 2018); an emerging format for publishing geotemporal data to the web, within dataset landing pages as an improved mechanism for publishing NASA datasets over the web. In particular WG Co-Chairs McGibbney and Armstrong proposed a strategy for achieving greater web visibility for NASA data sets based upon oceanographic coverage datasets archived within the NASA JPL's PO.DAAC (McGibbney & Armstrong, 2016). This initiative could be further extended to include use of a Linked Data concept such as (JSON-LD, 2018), providing important context for NASA geotemporal data on the web. Linked Data is about using the web to connect related data that wasn't previously linked or using the web to lower the barriers to linking data currently linked using other methods. More specifically, Wikipedia defines Linked Data as "*a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF.*" Through the provisioning of a CoverageJSON-LD context file, the objects and properties in a CoverageJSON file (e.g. sensor characteristics, data parameters, observed properties, units, ranges, calendars, coordinates systems, etc.) can be converted to URIs and triples and utilized in a linked data manner. This will enable integration with other linked data, possibly from interdisciplinary fields which promotes new uses for NASA science data. Additionally, the new CoverageJSON-LD structure and semantics would provide the ability to query datasets in ways which are currently impossible.

As of early 2019, the ESIP Semantic Technologies Committee are continuing work on an extension of schema.org (science-on-schema.org, 2019) which will provide concrete guidance on how Earth science-specific content can be encoded as structured data markup.

## 2.13 Recommendation 13: All NASA Websites should maintain a Sitemap to improve organization and prioritization of Website content

The Challenge

Currently, NASA DAACs do not provide key metrics used by commercial web crawlers to improve crawl strategy and hence improve interpretation of website content. Website dynamics representing each web page such as last modified date, change frequency, URL priority, etc. are not provided, hence they cannot be used within commercial search engine rankings.

The Recommendation Description

We recommend all NASA websites (especially those of DAACs) generate, maintain and offer sitemaps representing website content. Sitemaps are an easy way for webmasters to inform search engines about pages on their sites that are available for crawling. NASA website administrators can then submit their sitemaps directly to commercial search engine administrators as a mechanism for improving the way web crawlers navigate NASA websites.

Intended Outcome

A sitemap is a strategic first port of call for a web crawler to prioritize the way it navigates and traverses your website. Without one, a web crawler has no concept of 'priority', therefore the ranking and relevance scores need to somehow be created within the search engine itself as opposed to the web crawler providing a rich source of priority to the index scoring. Through provision of sitemaps, commercial web crawlers can harvest this information and feed it into search engine rankings.

Possible Approaches to Implementation

In its simplest form, a sitemap is typically an XML file that lists URLs for a site along with additional metadata about each URL (when it was last updated, how often it usually changes, and how important it is, relative to other URLs in the site) so that search engines can more intelligently crawl the site. The WG has made available, and continues to maintain Sitepod (Sitepod, 2018); a sitemap generator written in Hypertext PreProcessor (PHP) for addressing this recommendation. Sitepod is capable of generating sitemaps, in various encodings, for any given website e.g. NASA DAAC's. Sitepod also has the ability to post Sitemaps to commercial search engine providers such as Google and Yahoo, the purpose and intent being that web crawlers operated by these companies will be able to act and better interpret both the dynamism and importance of NASA websites and hence rank them accordingly within their search engine rankings. The WG hopes that Sitepod will be beneficial to all stakeholders who have an interest in promoting the characteristics of their website(s), collections, datasets and granules within the commercial search engine space.

## 2.14 Recommendation 14: Collect end user behavior of NASA search clients and infrastructure

The Challenge

By instrumenting digital tools which achieve data capture, we can understand human behavior and decision-making in ways never possible before. People rely on tools, such as DAAC websites and related software products, to consume news, obtain data, undertake scientific analyses, connect with others and generally *do* work. Collecting end user behavior will promote data discovery for the end user, while offering a method for the DAACs to facilitate discovery especially of

interdisciplinary datasets. Gathered metrics can be useful for multiple purposes and provide a check on the assumptions of those developing the search information and aggregation with data on the actual practices of the user community. This will improve tool usability, increase user adoption and enable improved understanding of the information and assets users make decisions with.

The Recommendation Description

We recommend that NASA search clients and interfaces should capture the full lifecycle of user behavior including web clicks, data downloads, and tools and services accessed.

Intended Outcome

The primary outcome of this recommendation is to provide dataset recommendations to end users, in a similar manner that eCommerce platforms such as eBay, Amazon, etc. offer supplementary information to end users. A well-recognized example is that user *A* selected *X* and therefore may also be interested in *Y and Z*. A secondary outcome however, will inevitably be significantly improved instrument/application logging, the ability to characterize user browsing/session interaction, rich visualizations which will enable NASA to understand not just *WHAT* people do, but also *HOW* people (stakeholders) work. When you understand how people work, you can improve how they work and the value they place on their tools. This has significant benefits for the entire ESDIS software architecture as well as individual DAAC software portfolios.

Possible Approach to Implementation

This recommendation makes a strong case for deriving a correlation between behavioral metrics and search rankings across ESDIS search infrastructure. With this in mind, let's consider Figure 5 which presents a typical basic behavioral browsing cycle for any given user. This could of course be any user of a DAAC website. Note, that a NASA-commissioned survey (Blink, 2017) found that when asked to find data across NASA DAAC's, the majority of users utilized commercial search engines such as Google as their primary search interface. This is represented by the use of the '*G'* icon at the top of Figure 5.

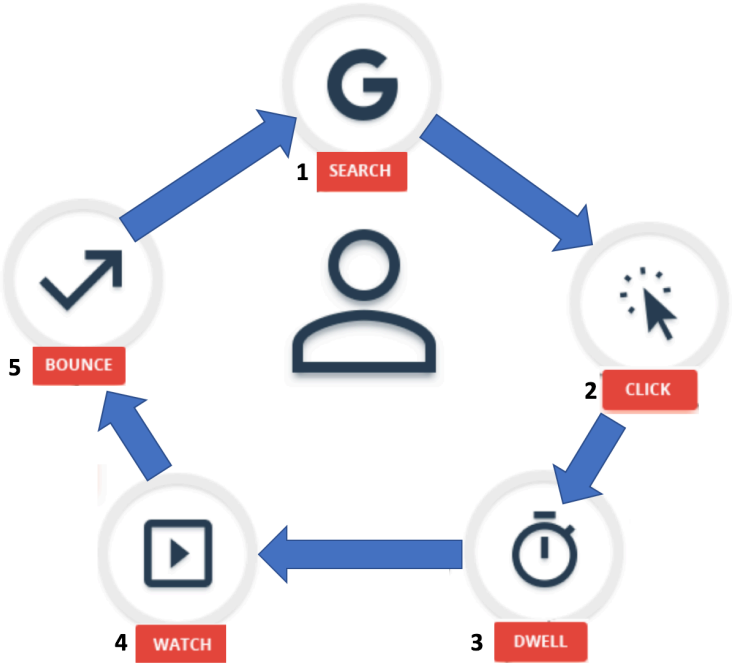


Figure 5: A typical basic behavioral browsing cycle

Figure 5 (interpreted from top in a clockwise manner) portrays the following key elements of the behavioral browsing workflow

1. Query execution; a user arrives at a search engine front page such as a DAAC homepage and then executes a query. Upon query execution, they are directed to a search engine results page (SERP) where they are presented with a ranked list of results from which they can expand their browsing session to (hopefully) relevant content.

2. Click through; the user typically selects one of the *top ten* results from the initial SERP which either links to internal or external content from the domain serving the SERP. It should be noted that at this stage, the user may not even know specifically what they are looking for yet e.g. they may not know which data or even dataset they are looking for. The total clickthrough stream generated at this stage of the user interaction may comprise the significant portion of a browsing session. Additionally, the content of a clock through stream is typically indicative of when a user is refining a selection to something highly relevant for their needs therefore modeling and understanding click streams is an integral aspect of the overall process.

3. Dwelling period; the user eventually finds content of a high enough interest that they spend *dwelling time* engaging with the content, possibly evaluating it for accuracy and/or appropriateness and determining whether or not it is of use or further relevance for their information retrieval purpose(s).

4. Watch/consume content; this is where a user would physically consume content e.g. engage in a data download from a DAAC site, watch a YouTube video, play a live news stream, etc. Until

now, this single element of the process workflow has been used as the primary metric for advancing usability analysis at ESDIS. Unfortunately, however, this metric alone does not provide the contextual information as to HOW people got to this stage and what their browsing behavior was until this point.

5. Return to results; once the user has, for example, downloaded the required data or finished the multimedia clip they were watching, they typically *bounce* back to the initial phase of the browsing cycle e.g. Phase 1 above. This either results in the execution of a new query or in many occasions returning to the SERP Phase 2 where the user engages in a new information retrieval scenario.

This recommendation advises that the above use case, and the behavior it portrays, be used to track user characterization as a start. In essence, this will enable search engine administrators to evaluate what a meaningful conversion rate is concerning the correlation between user browsing behavior and search relevance. A partial advance/success metric could initially be determined by the wealth of knowledge and insight provided by capturing the full life cycle of user behavior and correlating this with data downloads from NASA DAAC infrastructure. This will provide an initial, insightful and revolutionary approach to better understanding user needs while still maintaining user privacy and anonymity.

The WG effort which has fed into this recommendation is not mature enough to evaluate how combining user data from multiple EOSDIS systems e.g. CMR/ES/URS/EMS, can be used to analyze the datasets that are downloaded by the same users. This is recommended for future work which should be evaluated by future ESDSWG on a DAAC-by-DAAC basis.

## 3 Summary

This Technical Note presents the findings and recommendations of the 2015-2018 Search Relevance and User Characterization ESDS Working Groups. In condensed form the 14 recommendations derived from the topic subgroups are defined in the introductory section of this document. As of early 2019 this work has already had significant impact on core EOSDIS software infrastructure such as the CMR. This work has informed additional efforts taking place within the ESIP Semantic Technologies Committee regarding the advancement of an Earth science-specific extension (earthsci.schema.org) to the popular schema.org metadata initiative.

## 4 References

Allen, J. F. "Maintaining knowledge about temporal intervals". *Communications of the ACM* **26**(11) pp.832-843, Nov. 1983.

Apache Any23 – Anything to Triples (2019). Retrieved July 30, 2019 from https://any23.org

Apache OpenNLP (2019). Retrieved July 30, 2019 from http://opennlp.apache.org

Blink - Planet Centered Design – Transforming data into knowledge (2017). Retrieved July 30, 2019, from https://blinkux.app.box.com/s/bzfd2sa56sqgbriq50kk9zn03gr9n0bu

CoverageJSON - CoverageJSON (CovJSON) is a geospatial coverage interchange format based on JavaScript Object Notation (JSON). Retrieved July 30, 2019 from https://covjson.org/

Elasticsearch (2019) Retrieved July 30, 2019 from https://www.elastic.co/products/elasticsearch

Google - Structured Data Testing Tool (2019). Retrieved July 30, 2019 from https://search.google.com/structured-data/testing-tool

Google - Webmaster Search Console Tool (2019). Retrieved July 30, 2019 from https://www.google.com/webmasters/tools/home

Jarvelin, K. and Kekalainen J. (2000) IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41-48. ACM, 2000.

Jarvelin, K. and Kekalainen J. (2002) Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS), 20(4):422-446, 2002.

Jiang, Y., Li, Y., Yang, C., Liu, K., Armstrong, E., Huang, T., Moroni, D., & Finch, C. (2017): A comprehensive methodology for discovering semantic relationships among geospatial vocabularies using oceanographic data discovery as an example, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2017.1357819

JSON-LD - JSON for Linking Data (2019). Retrieved July 30, 2019 from http://json-ld.org/

Kilicoglu, H., Fiszman, M., Rodriguez, A., Shin, D., Ripple, A., & Rindflesch, T. C. (2008). Semantic MEDLINE: a web application for managing the results of PubMed Searches. Proceedings of the Third International Symposium for Semantic Mining in Biomedicine.

McGibbney, L. J., Armstrong, E. M. (2016) An Innovative Open Data-driven Approach for Improved Interpretation of Coverage Data at NASA JPL's PO.DAAC IN41B-1660, AGU Fall Meeting 2016, Dec 15th, 2016, San Francisco

Natural Language Toolkit (2019). Retrieved July 30, 2019 from http://www.nltk.org

Open Geospatial Consortium (2010). Implementation Standard for Geographic Information – Simple feature access – Part 2: SQL option, Version 1.2.1, Retrieved July 30, 2019 from http://portal.opengeospatial.org/files/?artifact_id=25354

Open Geospatial Consortium (2011). Implementation Standard for Geographic Information – Simple feature access – Part 1: Common architecture, Version 1.2.1, Retrieved July 30, 2019 from http://portal.opengeospatial.org/files/?artifact_id=25355

Raskin, R. (2005). Knowledge representation in the semantic web for Earth. Retrieved July 30, 2019 from http://www.sciencedirect.com/science/article/pii/S0098300405001020.

schema.org - Schema blog: Describing Datasets with schema.org (2012). Retrieved July 30, 2019 from http://blog.schema.org/2012/07/describing-datasets-with-schemaorg.html

schema.org - Dataset (2019). Retrieved July 30, 2019 from https://schema.org/Dataset

schema.org - DataCatalog (2019). Retrieved July 30, 2019 from https://schema.org/DataCatalog

science-on-schema.org (2019). Retrieved July 30, 2019 from https://github.com/ESIPFed/science-on-schema.org

Semantic MEDLINE - Semantic Knowledge Representation. (2012). Retrieved July 30, 2019, from http://skr3.nlm.nih.gov/SemMed

Sitepod (2019). Retrieved July 30, 2019, from https://github.com/nasa/sitepod

Stanford CoreNLP (2019). Retrieved July 30, 2019, from http://stanfordnlp.github.io/CoreNLP/

SWEET - Semantic Web for Earth and Environmental Terminology (2019). Retrieved July 30, 2019, from http://sweetontology.net

Yilmaz, E., Aslam, J.J., and Robertson, S. (2008). A new rank correlation coefficient for information retrieval. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08). ACM, New York, NY, USA, 587-594

## 5 Authors

### 2015-2018 Search Relevance Working Group

Ed Armstrong (Co-Chair), Lewis McGibbney (Co-Chair), Ross Bagwell, Nelson Casiano, Robert Downs, Peggy Eaton, Robert Ferraro, Lauren Frederick, Danielle Golon, Beth Huffer, Yongyao Jiang, Siri Jodha Khalsa, Stephan Klene, Jeanne Laurencelle, Christopher Lynnes, Megan Mitchell, James Norton, Steve Olding, Elli Pauli, Joshua Poore, Hampapuram Ramapriyan, Scott Saxon, Deborah Smith, Chris Stoner, TJM, Tammy Walker, Lalit Wanchoo, Stephanie Wasley, Vicky Wolf, Daine Wright, Yonsook Enloe, Grace Yang, Lindsay Spratt, Djoerd Hiemstra, Grant Ingersoll, Mihai Datcu, Ramkumar Aiyengar, Kyle Hundman, Kim Whitehall, Stefano Nativi, Nathan Clark, Daniel Di Silva

Dr. Lewis John McGibbney Ph.D., B.Sc.

Jet Propulsion Laboratory

California Institute of Technology

4800 Oak Grove Drive

Pasadena, California 91109-8099

Mail Stop: 158-256C

Email: lewis.j.mcgibbney@jpl.nasa.gov

Mr. Edward M Armstrong MS, BA

Jet Propulsion Laboratory

California Institute of Technology

4800 Oak Grove Drive

Pasadena, California 91109-8099

Email: edward.m.armstrong@jpl.nasa.gov

Edited by ESDIS Standards Office staff

eso-staff@lists.nasa.gov

## Appendix A - Glossary of Acronyms

| Acronym | Description |
|---|---|
| AOD | aerosol optical depth |
| ATBD | Algorithm Theoretical Basis Document |
| DAAC | Distributed Active Archive Center |
| DCG | Discounted Cumulative Gain |
| ECHO | Earth Observing System (EOS) Clearing House |
| EED | Evolution and Development |
| EMS | ESDIS Metrics System |
| EOSDIS | Earth Observation System Data and Information System |
| ES | Earthdata Search |
| ESDIS | Earth Science Data and Information System |
| ESDSWG | Earth Science Data System Working Group |
| FIRMS | Fire Information for Resource Management System |
| FS | Federated Search |
| GCMD | Global Change Master Directory |
| GSFC | Goddard Space Flight Center |
| IR | Information Retrieval |
| LARC | Langley Research Center |
| LPDAAC | Land Processes DAAC |
| nDCG | normalized discounted cumulative gain |

| | |
|---|---|
| NLP | Natural Language Processing |
| NSIDC | National Snow and Ice Data Center |
| OGC | Open Geospatial Consortium |
| PO.DAAC | Physical Oceanography DAAC |
| RTS | Rich Text Snippets |
| SIPS | Science Investigator-led Processing Systems |
| SWEET | Semantic Web for Earth and Environmental Terminology |
| UMLS | Unified Medical Language System |
| URL | Uniform Resource Locator |
| URS | User Registration System |