RESEARCH PAPER

Improving NASA's Earth Satellite and Model Data Discoverability for Interdisciplinary Research, Applications, and Education

ZHONG LIU
CHUNG-LIN SHIE
SUHUNG SHEN
JAMES ACKER
ANGELA LI
JENNIFER C. WEI
DAVID J. MEYER

*Author affiliations can be found in the back matter of this article

ABSTRACT

Since the Internet era began, numerous earth science data services have been developed to facilitate data discovery (e.g., data sources, documents, facts, visualization, opinions) and data access for research and application activities. For example, a large collection of NASA's earth science data has been made searchable and freely downloadable over the Internet. Some value-added services even allow users to analyze and visualize many variables online (e.g., 2,000+ in NASA Giovanni) without downloading data and software.

However, finding and discovering suitable datasets and information for interdisciplinary research (involving two or more scientific disciplines), applications, education, and other emerging activities (e.g., water, food, energy nexus) has been a challenge not only for users, especially for those who are unfamiliar with scientific disciplines, measurements, or models, but also for data producers and service developers who want their data to be discoverable. NASA earth science data are currently archived and distributed at twelve NASA discipline-oriented Distributed Active Archive Centers (DAACs). Even though most datasets are online, search results often contain many similar datasets with limited information for self-guided or heuristic dataset discovery, so some users simply send an inquiry to DAAC support staff for advice. Conducting interdisciplinary research often requires multiple datasets from different data repositories, which can make it even harder to find and discover suitable datasets without additional data services to accommodate these emerging user needs.

In this article, we assess current data discovery practices and publications (e.g., reports from working groups) to identify challenges and make actionable recommendations for improving earth science data discoverability and facilitating interdisciplinary activities.

Highlights:

- Status review of earth science data discoverability for interdisciplinary research and applications at NASA GES DISC.
- Review of data discovery research and working group activity.
- Discuss challenges and opportunities for interdisciplinary data discovery with recommendations for practitioners.

CORRESPONDING AUTHOR:

Zhong Liu

AuthorAffilation

Zhong.Liu@nasa.gov

KEYWORDS:

data discoverability; NASA; earth science; remote sensing; interdisciplinary; recommendations

TO CITE THIS ARTICLE:

Liu, Z, Shie, C-L, Shen, S, Acker, J, Li, A, Wei, JC and Meyer, DJ. 2023. Improving NASA's Earth Satellite and Model Data Discoverability for Interdisciplinary Research, Applications, and Education. *Data Science Journal*, 22: 9, pp. 1–16. DOI: https://doi.org/10.5334/dsj-2023-009

1. INTRODUCTION

Scientific data discovery (datasets, documents, facts, visualization, opinions) (e.g., Weikum 2013) often requires users to possess sufficient scientific knowledge to pose useful search questions, along with tools allowing data service providers to be able to correctly understand what users search for and provide usable search results, especially when users search for unfamiliar datasets or information content. On the other hand, data services can also enable self-guided search with capabilities (e.g., spatiotemporal bounding) and abundant dataset-related information (e.g., publications, user forums). Enabling data discovery is listed as one of the challenges (Behnke et al. 2019) for NASA's Earth Observing System Data and Information System (EOSDIS) (NASA EOSDIS 2022a), which manages 12 discipline-oriented NASA Distributed Active Archive Centers (DAACs) (NASA DAACs 2022). In a recent FAIR (findable, accessible, interoperable, and reusable) (Wilkinson et al. 2016) data assessment for all NASA DAACs (Ramapriyan & Behnke 2020), data findability received the lowest score among all four FAIR categories. In short, data discovery has been a challenge not only for users but also for data producers and data service providers who want their data to be effectively discoverable for maximizing their data distribution.

In this era of rapidly increasing data availability, finding suitable datasets for research, applications, education, and other emerging activities (e.g., water, food, energy nexus) has become increasingly challenging. This is especially true for those who are unfamiliar with scientific disciplines, measurements, or models. At present, datasets are largely archived and disseminated based on data types (e.g., satellite retrievals, field campaigns, models) or disciplines (e.g., each of the 12 NASA DAACs that specialize in certain or multiple disciplines (NASA DAACs 2022)).

Finding data for interdisciplinary activities (involving two or more scientific disciplines; e.g., agriculture and water management) is even more challenging because users often need to visit multiple discipline-oriented data archives and have adequate information to identify suitable datasets. This can be particularly difficult for inexperienced users or users who are not familiar with datasets from other disciplines. The lack of uniform user experiences with different data repositories is another challenge to the search for suitable data products. Currently, most data services or tools are developed for certain groups of scientists or principal investigators in their special disciplinary community. As a result, data service developers often do not have enough knowledge to design data services accommodating users in other disciplines. In addition, different vocabularies used in different communities (e.g., Parsons et al. 2022) can confuse users attempting to both explore and use various data services.

The FAIR data guiding principles (Wilkinson et al. 2016) start with *findable*, which is (or may be) one of the most challenging tasks for data users, data producers, and service providers worldwide. In the Internet era, data services that facilitate data discovery heavily rely on metadata provided by data producers (e.g., Bugbee et al. 2021; Mathiak et al. 2023). However, many data producers are neither aware of, nor paying attention to, the importance of including sufficient and standardized metadata in their datasets for improving data discovery and usage, mainly because best community data practices and standards have often not been required in their research proposals, such as preparing a data management plan. As a result, data products often do not include sufficient and standardized metadata. This situation has made data service development difficult. For data service developers, without sufficient and standardized metadata from data providers, extra efforts are needed to add additional metadata, which can be a difficult task for many data repositories due to the lack of adequate disciplinary knowledge among staff and the amount of work involved. Furthermore, without proper metadata in a dataset, users will need to seek additional relevant resources (e.g., product documents, research publications), which are often either missing or insufficient.

Despite the reality that most datasets are online with certain heuristic search capabilities available (e.g., filtering), search results often contain many similar datasets that are designed for various research or application purposes by different projects or missions. Without additional information (e.g., publications, usage examples, FAQs), it is often difficult for users to conduct self-guided data discovery (Mathiak et al. 2023). As a result, some users simply ask data support staff for advice.

Lacking unavailable data products that users want is another challenge for both data producers and service providers (e.g., Wu et al. 2019). For example, when a user tries to look for a daily precipitation dataset from a repository but only data with half-hourly or hourly temporal resolution is available, the user may receive a 'no daily data found' result. If the daily precipitation dataset is also being provided by the product or service provider as a value-added product, the user will have less difficulty finding the dataset they want or need.

Earth science data users have diverse backgrounds—researchers, application users, educators, students, and ordinary citizens—and possess different knowledge and expertise in handling data and information. Meeting their diverse needs suggests that developing and providing different services (e.g., user interfaces and information contents) is necessary to facilitate data discovery and access. For instance, for a person who simply wants an annual total precipitation map for their region of interest, a traditional workflow for users (e.g., online finding and downloading data followed by processing and visualizing the dataset offline) may not fit this person's quick need. A more efficient tool, such as Giovanni (NASA Giovanni 2022; Liu and Acker 2017; Acker and Leptoukh 2007), can provide the set of needed procedures (i.e., finding, processing, and visualizing, all online, without having to download data and software) and efficiently (properly and quickly) produce the result (plotting) that fulfills a user's specific needs (which we would like to provide users with a 'Window Shopping' service). In short, meeting users' needs also plays an imperative role in designing data services that facilitate data and information discovery.

Over the years, there have been several research community activities that have produced recommendations for improving data discovery (e.g., Contaxis et al. 2022). For example, Wu et al. (2019) collected and analyzed 79 data discovery/search scenarios and developed 10 recommendations for data service developers. Another example is the recommendations for earth science data search relevance developed by McGibbney et al. (2019) as a part of NASA's Earth Science Data System Working Group (ESDSWG) activity. Some discipline-specific websites also post information that guides users on how to select data (e.g., Huffman 2022; NASA Earthdata 2022a, 2022b). As previously mentioned, many existing data services have been developed for a particular dataset or discipline. Therefore, those recommendations may not be suitable for data services for interdisciplinary research and applications in which datasets are involved from several disciplines and archived at different data repositories. The current situation warrants further investigation into this new challenge by reviewing current practices in order to provide practical recommendations.

In this paper, we assess current operational practices in data discovery and publications (e.g., referral papers, reports from working groups). As one of the DAACs managed by NASA's EOSDIS (NASA EOSDIS 2022a), the NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) (NASA GES DISC 2022a) has archived and distributed multidisciplinary satellite and model data products. Although GES DISC only archives a portion of NASA earth data and users may also need data from other DAACs for their activities, the diverse and interdisciplinary data collection at GES DISC can still serve as an example or use case for this study. Based on the findings of this study, we discuss challenges and opportunities for improving earth science data discoverability and facilitating interdisciplinary research and applications. At the end, we provide practical recommendations. These recommendations may not be limited to GES DISC or NASA.

The structure of the paper is as follows: section 2 overviews existing operational practices, section 3 includes a summary of referral publications and reports from working groups, section 4 discusses challenges and opportunities, and section 5 provides our summary and recommendations.

2. DATA DISCOVERY PRACTICES AT NASA GES DISC

Established in the mid-1980s, NASA GES DISC (NASA GES DISC 2022a) is in Greenbelt, Maryland. It currently archives a total data volume of 3.4 petabytes consisting of 150 million data files and covering over 3,000 public and restricted multidisciplinary data collections, including atmospheric composition, water & energy cycles, climate variability, carbon cycle & ecosystem from both major NASA satellite missions (e.g., Global Precipitation Measurement (GPM)) and projects (e.g., the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2)). MERRA-2 provides NASA's global atmospheric reanalysis from 1980 onward. Enhancements have been made in MERRA-2, including the use of an upgraded version of the Goddard Earth Observing System Model, Version 5 (GEOS-5), data assimilation system; updates to the model (Molod et al. 2012; Molod et al. 2014) and to the Global Statistical Interpolation (GSI) analysis scheme (Wu et al. 2002); the first global reanalysis to assimilate space-based observations of aerosols and their interactions with other physical processes in the climate system; and a representation of ice sheets over Greenland and Antarctica (Bosilovich et al. 2016). In short, significant steps toward NASA's Earth system reanalysis goal have been taken in MERRA-2.

The GES DISC provides data services and support to users around the world, including (1) metadata support, documentation, and metrics (Liu et al. 2022) for archived datasets; (2) web-based discovery and access to data products and data download; (3) value-added services on data; (4) user services providing support for data access and use; and (5) community engagement and outreach (e.g., user working groups, workshops, trainings, conferences, webinars).

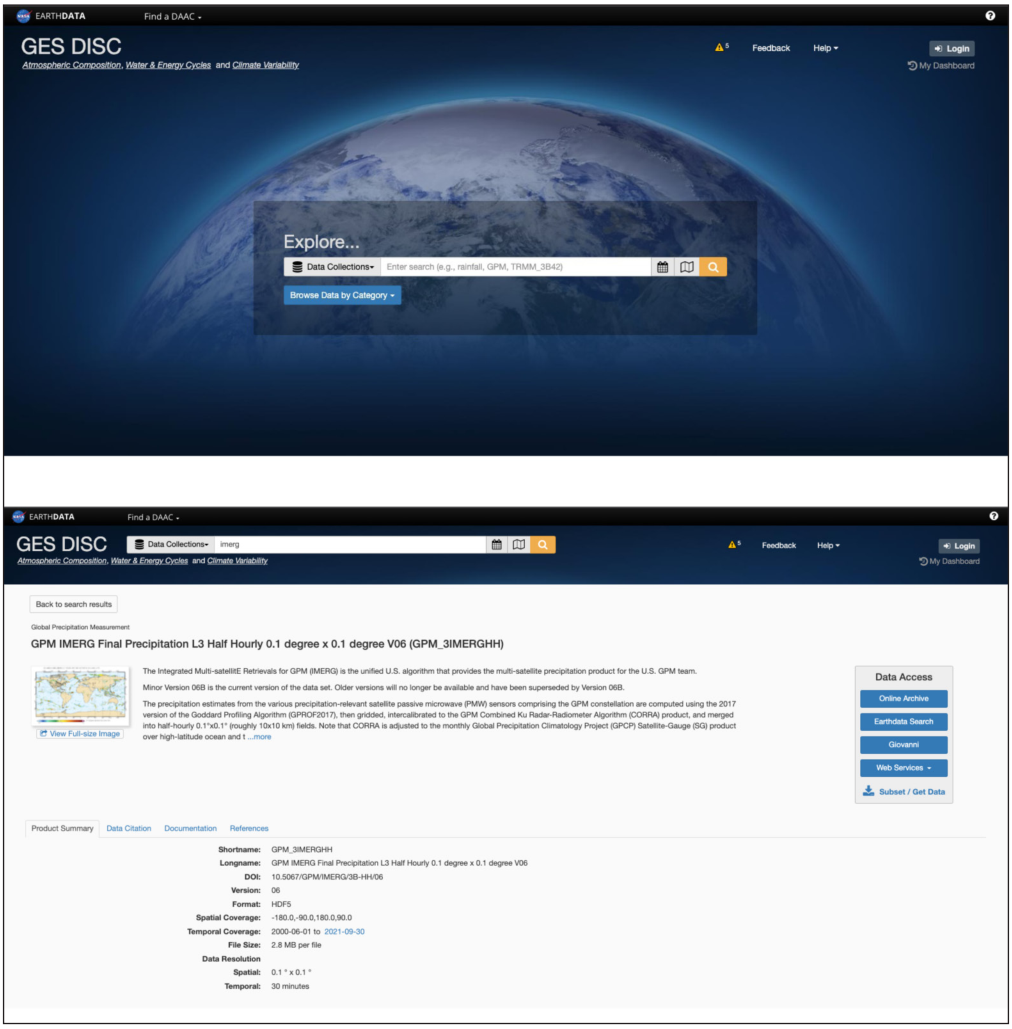
Over the years, data services at GES DISC have continuously evolved. Guided by these user support activities, best practices, and recommendations from workshops and publications, data service needs have been routinely identified and prioritized based on several criteria, such as available resources, level of difficulty, and user needs. A group consisting of scientists and software developers is formed to formulate implementation details (e.g., service requirements, user interfaces) and acceptance criteria. Metrics (Liu et al. 2022) are routinely collected and evaluated for service improvements.

For example, in the past, users could only download data in the original forms provided by data producers. Since the spatial coverage of most NASA datasets is global, users would have to download the entire global dataset even for local or regional studies or applications if (or when) data subsetting services were not available. This action would increase network congestion and data server loads and cause unnecessary data downloads. Today, GES DISC provides a range of subsetting capabilities, such as spatial subsetting, as well as regridding and reformatting services. With the ongoing cloud evolution (NASA Earthdata 2022a, 2022c; NASA GES DISC 2022b), data services will be significantly improved, especially for handling voluminous global satellite and model datasets that are difficult and inefficient for on-premises services to handle. In addition, users will no longer have to visit multiple DAACs for data services (e.g., download multidisciplinary datasets). In short, cloud environments will provide a wide range of data services on one platform that are currently difficult to provide on-premises.

The GES DISC web portal (Figure 1) (top) provides a Google-like interface for searching data and information (e.g., documents, how-to recipes, FAQs). Prior to this, users had to visit multiple GES DISC websites for needed data services and information, which could be confusing, inconvenient, and time-consuming, as well as difficult for them to remember those websites. The current GES DISC web portal (NASA GES DISC 2022a) has unified these data service websites to provide a one-stop shop for all data-related services and information. In Figure 1 (bottom), each dataset has its own dataset landing page (DLP) including such information as product summary, data access, data citation, and supporting documentation.

Current self-guided search methods are limited to keyword search (NASA GES DISC 2022a). General rules include a single keyword (e.g., product short name, platform short name, measurement, project name), multiple keywords, and simple query string operators (e.g., AND, OR, exclusion, and wildcard) (NASA GES DISC 2022a) for multiple keywords. Advanced search options include spatial and temporal range refinements. Search capabilities are still being developed in terms of relevance and accuracy.

NASA's Common Metadata Repository (CMR) (NASA Earthdata 2022d, Bugbee et al. 2021) is the backend engine behind GES DISC data search and other NASA data services, such as Earthdata (NASA Earthdata 2022a). CMR is a high-performance, high-quality, and continuously evolving metadata system. With CMR, all data and service metadata records for NASA's EOSDIS system are cataloged. CMR is also the authoritative management system for all EOSDIS data, including those at GES DISC and other DAACs. Metadata in CMR provide the description of a dataset; therefore, the quantity and quality of metadata records play an imperative role in data search and discovery (Bugbee et al. 2021).



Global Precipitation Measurement

GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06 (GPM_3IMERGHH)

The Integrated Multi-satellitE Retrievals for GPM (IMERG) is the unified U.S. algorithm that provides the multi-satellite precipitation product for the U.S. GPM team.

Minor Version 06B is the current version of the data set. Older versions will no longer be available and have been superseded by Version 06B.

The precipitation estimates from the various precipitation-relevant satellite passive microwave (PMW) sensors comprising the GPM constellation are computed using the 2017 version of the Goddard Profiling Algorithm (GPROF2017), then gridded, intercalibrated to the GPM Combined Ku Radar-Radiometer Algorithm (CORRA) product, and merged into half-hourly 0.1°x0.1° (roughly 10x10 km) fields. Note that CORRA is adjusted to the monthly Global Precipitation Climatology Project (GPCP) Satellite-Gauge (SG) product over high-latitude ocean and t ...more

Shortname:	GPM_3IMERGHH
Longname:	GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06
DOI:	10.5067/GPM/IMERG/3B-HH/06
Version:	06
Format:	HDF5
Spatial Coverage:	-180.0,-90.0,180.0,90.0
Temporal Coverage:	2000-06-01 to 2021-09-30
File Size:	2.8 MB per file
Data Resolution	
Spatial:	0.1 ° x 0.1 °
Temporal:	30 minutes

Figure 1 Top: The GES DISC web portal provides a Google-like interface for searching data and information (e.g., documents, how-to documents, FAQs). Bottom: Each dataset has its own dataset landing page (DLP) that includes complete information for product summary, data access, data citation, and documentation, respectively.

Search results from the GES DISC web portal (NASA GES DISC 2022a) can be refined by subject, measurement, source, processing level, project, and spatiotemporal resolution. However, finding a specific variable can still be a challenge. Most data products at GES DISC are packed as data collections; for example, the data collection of time-averaged two-dimensional monthly means (M2TMNXFLX) in MERRA-2 contains a total of 46 variables. If a user wants one specific variable (e.g., total precipitation) in this collection, the user must find the data collection first and then use the subsetting service or the dataset document to identify the variable, which can be difficult for users who are unfamiliar with the collection name and the DLP information.

Users or individuals with knowledge of dataset names can usually narrow down their search by using the dataset short names. However, only a small percentage of users (e.g., in Figure 2) may be familiar with dataset short names. For example, the Integrated Multi-satellitE Retrievals for GPM (IMERG) is a very popular global precipitation data suite that provides merged multisensor and multisatellite global precipitation estimates ranging from 30 minutes and daily to monthly. When a user searches for 'IMERG monthly' for the monthly IMERG dataset, the search results will consist of over 500 datasets. By contrast, a search for the short name, GPM_3IMERGM_06, or 'IMERG + monthly,' will only return one result that links to the exact DLP wanted and/or needed. These two keyword searches are, however, not intuitive.

To better understand a user's search habits, Google Analytics (Liu et al. 2022) is used. Figure 2 (top) shows that the default search type for the Google-like search interface (Figure 1) is the most searched content type. The most popular dataset keyword searches shown in Figure 2 (middle) are associated with satellite precipitation, hydrology, and atmospheric data assimilations (e.g., Global Land Data Assimilation System (GLDAS)), which is not a surprise because GES DISC is home to data from several major NASA precipitation measurement missions (e.g., Tropical Rainfall Measuring Mission (TRMM), GPM) and projects (e.g., Global Precipitation Climatology Project (GPCP), North American Land Data Assimilation System (NLDAS), MERRA-2).

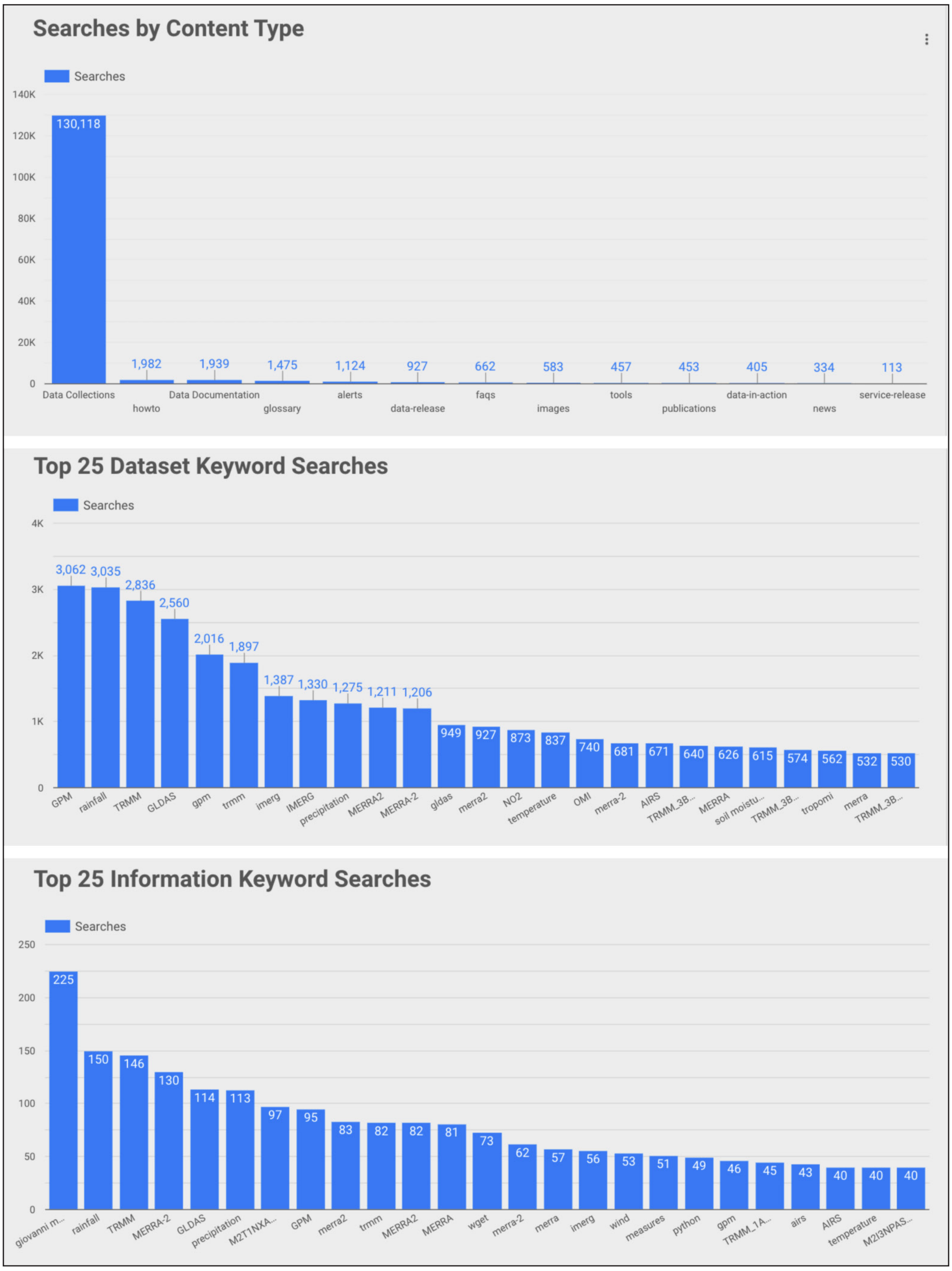


Figure 2 Metrics collected from Google Analytics (October 1, 2021–September 30, 2022). Top: Searches by content type. Middle: Top 25 dataset keyword searches. Bottom: Top 25 information keyword searches. 'Giovanni Measurements' tops the list.

The top searched keyword for information (e.g., FAQs, data documentation) is 'Giovanni Measurements.' As previously mentioned, Giovanni is a popular and powerful online tool developed at GES DISC. At present, there are over 2,000 variables served in Giovanni (NASA Giovanni 2022). In the GES DISC web portal, users can only search datasets and not their variable contents. In Giovanni, users or individuals can directly search for data variables (e.g., precipitation, temperature) that are likely more familiar to them, simplifying and expediting data discovery and access. Giovanni can be used for data evaluation, intercomparison, and other activities without downloading data and software (NASA Giovanni 2022; Liu and Acker 2017; Acker and Leptoukh 2007).

Recognizing the limitations (e.g., single dataset-oriented services) in data services and challenges for interdisciplinary data discovery, GES DISC has previously experimented with a novel search relevance method by allowing in-house data specialists to group related and frequently used datasets by research or application subjects (e.g., agriculture, hurricane). A keyword search for these subjects will return the associated datasets properly (NASA GES DISC 2022a, 2022c). However, this approach can be subjective, and the compiled variables may not be suitable for all research activities.

Even with a smaller list of search results, it is still difficult to find a suitable dataset. Previous work and external research publications play an important role in providing additional information (e.g., examples or use cases) (Lafia et al. 2016). Users can use these past investigations available on the DLP to learn how each dataset is utilized in research or applications. An ongoing novel activity at GES DISC is to use AI/ML to harvest science-subject-based use cases from published journal articles (Stoyanova et al. 2021).

Once a dataset is identified in the search results, the user will be directed to its DLP (Figure 1). At present, each DLP (Figure 1) includes key information about a dataset, such as data summary, data access, citation, documentation, and references. The concept of DLP is not novel and is widely used in product-related services like Amazon because it provides a one-stop shop for all dataset-related services and information. The DLP continues to evolve with improvements. More dataset-related information and services need to be added or expanded (e.g., publications). Adding a user forum can also provide useful feedback for product developers to identify product issues. It can help data centers in assisting new users to get help from other experienced users, like services provided on current commercial shopping sites.

A significant number of scientists, including data specialists at GES DISC, are not trained to deal with multidisciplinary science subjects and datasets. Therefore, putting together a list of interdisciplinary datasets can be a challenge, especially if/when it is strongly dependent on the knowledge of data specialists who are often familiar with a few satellite missions or projects but may not be aware of similar datasets from other missions or projects. Ideally, data producers should provide such information (e.g., data usage) in their product metadata for data services, but again, as mentioned, there is no mandate to include such information in the current data management plan for a data producer.

GES DISC is also developing personalized data services in 'My Dashboard' for registered users (NASA GES DISC 2022a). With the dashboard, registered users can bookmark their favorite DLPs and information, automatically record site (visit and access) history, and share those links with other colleagues. Users can also manage the dashboard for activities, such as importing and exporting links. With such personalized services, users do not need to repeat the same search and discovery processes each time they visit.

3. REVIEW OF RESEARCH AND WORKING GROUP RECOMMENDATIONS

3.1 RESEARCH PROGRESS

Over the years (since the Internet era) several research activities have been conducted to better understand data discovery challenges and provide recommendations for practitioners and project management. For example, after assessing the status of data discovery, Weikum (2013) identified a gap in commercial search engines that can only satisfy popular information needs by typical users, as opposed to expert needs by advanced users. Weikum (2013) concluded that several key capabilities (i.e., search, discover, compile, and analyze relevant information) play an important role in satisfying a user's specific task. Weikum (2013) presented a 10-year vision in which users will be able to conduct semantic search and information discovery, other than applying keywords and visiting pages. A few use cases were presented, such as science, humanities, business, and media analysts, among others. One key challenge is to semantically understand user search contents and be able to extract what users want. Weikum (2013) gave three recommendations: (1) knowledge search capabilities; (2) personalization and sociocultural awareness as a part of the capabilities; and (3) federated services to connect different components. A list of research directions based on other research activities was compiled, ranging from 'searching for knowledge' to user interfaces (Weikum 2013).

Several other recommendations have been proposed by different researchers. For example, Wu et al. (2019) collected and analyzed 79 data discovery use cases. After applying usability heuristic evaluation and expert review methods, they developed 10 recommendations for service developers at data repositories to consider for improving data discoverability and user experiences in data search. These recommendations can be summarized as providing (1) multiple ways (e.g., interfaces) to find data; (2) easy-to-read information (e.g., metadata, references, data usage metrics); and (3) consistency with other data repositories (e.g., standards). Another example is the 10 simple rules for improving research data discovery (Contaxis et al. 2022). In addition to providing thoughtful and rich information (e.g., metadata, publications), Contaxis et al. (2022) added additional rules for the level of data access and ethical standards.

There have been several research and application activities regarding the semantic web for earth and environmental sciences (e.g., Raskin & Pan 2005; Li et al. 2014; Fox et al. 2015; Wang et al. 2018; Wang et al. 2023). A few articles were included in the e-book (Narock & Fox 2015) *The Semantic Web in Earth and Space Science: Current Status and Future Directions*, outlining the current state of the field, emerging challenges, and future directions using mature semantic applications within the geosciences. Semantic websites rely on vocabularies and ontologies to classify and explain entities. Examples of semantic websites (e.g., Google, Best Buy) use a vocabulary to associate meaning with data on the web (Devopedia 2022). The vocabulary is defined by the community. It is not an easy task to develop such a vocabulary for interdisciplinary research in which vocabularies can be different among disciplines (e.g., Parsons et al. 2022).

3.2 WORKING GROUP RECOMMENDATIONS

In 2015, the NASA ESDSWG (McGibbney et al. 2019) was formed to develop search relevance recommendations for data service development in the NASA earth science data service community (e.g., the 12 DAACs). The working group delivered 14 recommendations (McGibbney et al. 2019) that cover the following topics: (1) spatiotemporal relevance; (2) dataset relevance heuristics; (3) semantic dataset relationships; (4) federated search; (5) utilization of commercial search engines; and (6) user characterization. Compared to other recommendations previously mentioned, these recommendations provide more practical directions for implementation, such as spatiotemporal relevance. In a full comparison, there are overlapping areas of similarity evident in these recommendations, such as utilization of dataset-related information (e.g., metadata, metrics), personalization, and additional actionable items (e.g., spatiotemporal relevance).

Several other groups from domestic and international organizations have been working on data discovery challenges (e.g., ESIP 2022a, RDA 2022a). The Earth Science Information Partners (ESIP) (ESIP 2022b) community is a group of data and information technology practitioners. ESIP provides many collaboration areas or clusters that are made up of administrative committees and small working groups where participants from different agencies or organizations (e.g., NASA, NOAA) work together and tackle challenges. One of them is the discovery cluster (ESIP 2022a). GES DISC has implemented some activities, including linking datasets that are used for ESIP (2022a). The Data Discovery Paradigms Interest Group (RDA 2022a) in the Research Data Alliance (RDA 2022b) is another group for improving data discovery. The goal of the group is to develop guidelines and recommendations that can be adopted by data repositories. Activities of the group are also related to those of ESIP and NASA (RDA 2022b). Best practices are being drafted for data providers, repositories, and data seekers, respectively. Most practices are consistent with previously published work, but special needs for interdisciplinary activities have not been adequately addressed yet.

3.3 OTHER ACTIVITIES

To meet user needs, some disciplinary organizations have put together helpful information pages to guide users to select datasets. For example, there is an introduction to global precipitation algorithms and datasets, written by Huffman (2022), available on the website of the International Precipitation Working Group (IPWG 2022). Huffman (2022) provided a background and descriptions of major algorithms and datasets, which could help new users to select a suitable precipitation dataset. In addition, NASA Earthdata (NASA Earthdata 2022b) develops data pathfinders, a guide that provides a brief introduction to the data, use cases, other resources, and the benefits and shortcomings of remote sensing data for several interdisciplinary subjects, such as farming and water resources, disasters, and disease transmission.

4. DISCUSSION OF CHALLENGES AND IMPROVEMENTS

Over the years, efforts have been made to improve data discovery by involving data repository practitioners, researchers, and working groups to collaborate on this activity. Recommendations have either been implemented or are being prototyped, as seen in the evolution of GES DISC data services. However, there are still numerous improvements to be made in data discoverability not only for a single dataset but also for multiple datasets. These datasets are often used in interdisciplinary research and applications.

Improving data discoverability involves many factors. Over the years, many rules and recommendations have been developed in previous research and working group reports, presenting different degrees of difficulty in implementation. Some of them (e.g., incomplete metadata and information, lacking standard compliance, federated search) need additional community-level efforts, which may exceed the scope of an individual data repository. For the time being, a more feasible way for data repositories is to further enhance their heuristic capabilities (e.g., providing additional dataset-related information and linking relevant datasets). The following discussion will focus on implementation feasibility.

4.1 BETTER UNDERSTANDING OF USER INQUIRIES

It has been almost 10 years since Weikum (2013) developed a 10-year vision for a quantum leap in services (e.g., semantic search) that would meet advanced user needs. Although there have been several research activities (e.g., Li et al. 2014; Huffer et al. 2015; Augustin et al. 2019), there is still a gap between research and operation. For nonprofessional users, finding data services is equally as challenging as finding data. According to the ACSI survey (ACSI 2022; NASA EOSDIS 2022b), nonprofessional users have been giving the lowest satisfaction scores to data services provided by 12 NASA DAACs since the survey began.

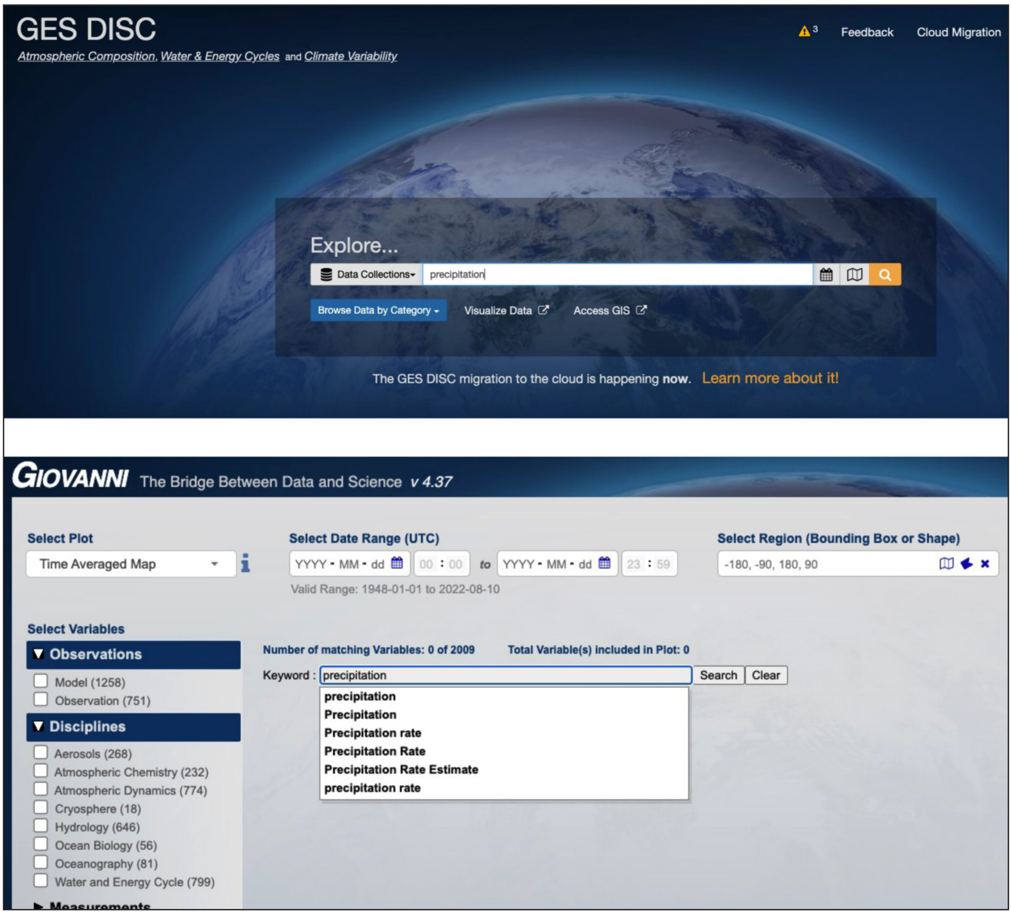
Currently, unless a dataset DOI (digital object identifier) or a link to a DLP is known, most GES DISC users depend on the Google-like search interface in the GES DISC web portal or commercial search engines to find data and information, as seen from the data services metrics at GES DISC. Small (e.g., with only a few datasets) data repositories normally provide a list view of their products and do not provide search interfaces because they are simply not needed.

Understanding user inquiries correctly plays a key role in data discoverability (Weikum 2013), which could be a part of the reason that progress has been slow in semantic search research and applications. In most cases, search terms or email messages sent to data repository support staff are vague (e.g., 'precipitation,' 'temperature'). Without additional information or interactions, users either retry with different search terms or use other means (e.g., spatiotemporal resolution) to refine search results. This situation will continue even when natural language processing (NLP) is implemented. Several iterations are often needed in data services (e.g., user interfaces) to improve the understanding of user inquiries, which may need further research and prototyping experiments.

Currently, the most feasible way to improve understanding of user inquiries is to enhance heuristic search capabilities. Search suggestions have increasingly gained popularity in many search engines, such as Google. Adding a drop-down list of search suggestions and refining these discernment capabilities can be very helpful to users or individuals. For example, on the GES DISC main page, when one searches for 'precipitation,' there is no additional suggestion (Figure 3); by contrast, in GES DISC Giovanni, when one searches for 'precipitation,' a list of suggestions (e.g., precipitation rate, precipitation rate estimate) is presented (Figure 3), which could be helpful for interdisciplinary activities as well because precipitation can have alternate nomenclature in related disciplines. To implement search suggestions efficiently, they need to change the current search results, that is, from data collection to variable or parameter, which is described next. Also, search suggestions highly depend on metadata, which is often missing or insufficient in datasets. Staff at data repositories can help add additional metadata.

4.2 DATASET COLLECTIONS

One of the areas of improvement is data presentation. As previously mentioned, most data products at GES DISC are packaged as data collections, and finding a variable can be difficult. Therefore, data search could be improved by switching from collection to variable. Adding a tab in the DLP (Figure 1) showing a list of variables can be an immediate improvement. A successful example is found in Giovanni, where users can search over 2,000 variables with keywords they are familiar with, such as 'precipitation.' In Giovanni, users can find variable names from different disciplines, which may help them find the variables they are familiar with.



Data Collections: precipitation

The GES DISC migration to the cloud is happening **now**. Learn more about it!

Select Plot: Time Averaged Map

Select Date Range (UTC): YYYY - MM - dd 00 : 00 to YYYY - MM - dd 23 : 59

Valid Range: 1948-01-01 to 2022-08-10

Select Region (Bounding Box or Shape): -180, -90, 180, 90

Select Variables

Observations

- □ Model (1258)
- □ Observation (751)

Disciplines

- □ Aerosols (268)
- □ Atmospheric Chemistry (232)
- □ Atmospheric Dynamics (774)
- □ Cryosphere (18)
- □ Hydrology (646)
- □ Ocean Biology (56)
- □ Oceanography (81)
- □ Water and Energy Cycle (799)

Number of matching Variables: 0 of 2009 Total Variable(s) included in Plot: 0

Keyword : precipitation

- precipitation
- Precipitation
- Precipitation rate
- Precipitation Rate
- Precipitation Rate Estimate
- precipitation rate

Figure 3 Top: A keyword search for 'precipitation' in the GES DISC web portal. Bottom: A keyword search for 'precipitation' in Giovanni, showing a list of suggestions.

Furthermore, GES DISC staff, as aforementioned, can add more metadata to suggestions (e.g., droughts, floods, agriculture, water management) to enhance search capabilities.

As NASA's Unified Metadata Model (UMM) (NASA Earthdata 2022e) rolls out, data search will be significantly improved at the variable level. The UMM is an extensible metadata model that provides a crosswalk for mapping between CMR-supported metadata standards (NASA Earthdata 2022e). The UMM includes EOSDIS concepts that include collections, granules, services, variables, visualizations, tools, and elements common to multiple UMM component models. In particular, the UMM-Var (NASA Earthdata 2022e) provides metadata about variables in EOSDIS data products, which plays a crucial role in the development of variable-oriented data services, such as data search, which requires such variable information. However, different vocabularies in different disciplines (Parsons et al. 2022) can be a challenge for data services to support interdisciplinary activities. Given the importance of dataset-related information (e.g., satellite anomalies, data usage, limitations) in self-guided data discovery, the UMM needs to add UMM-Information to facilitate data discovery.

In addition to metadata improvement, curated data collections at GES DISC and NASA Earthdata Pathfinders (NASA Earthdata 2022b) can partially bridge the gap between single and multiple dataset searches. One potential drawback for collections discovered via the pathfinders is that it can miss other relevant variables. For example, in the Agricultural and Water Resources in the Data Pathfinders, other potentially useful precipitation datasets can be missed, such as GPCP, which provides a long-term, carefully calibrated, and consistent climate data record available from 1983 to the present.

4.3 USER INTERFACES

User interfaces (UI) are the gateway to data search. A typical data search UI consists of three main components: a text search box, a calendar for time range, and a global map for spatial range. At GES DISC, a list of search categories (Figure 1), such as data documentation, FAQs, news, and tools, is provided to help users search for data and information. By combining categories with filtering capabilities, the implementation of spatiotemporal range can be very useful for users. For example, although there are thousands of data collections available at GES

DISC, it should be readily easy to find datasets for weather-related case studies, such as the weather conditions for the tragic Air France 447 crash that claimed 228 passengers and crew on board (Wikipedia 2022). If the search ranges are available in the UI, datasets that are outside the user-defined spatial and temporal ranges are not included in the search results, and only a few are relevant for this case study (e.g., the NCEP CPC global merged IR dataset available at a 30-minute interval, or MERRA-2 reanalysis). Although the spatial and temporal ranges are not fully implemented, they are included in some level-2 dataset subsetting services at GES DISC, and users can specify a point, a circle, or a rectangular box to search and subset data.

Additional UI improvements include the addition of shapefile capabilities. For example, in Giovanni, shapefiles such as countries, US states, land/sea masks, major watersheds, and large lakes have been included. Likewise, shapefiles can be added to spatial ranges in the GES DISC data search UI to facilitate dataset search. Event studies (e.g., floods) can be a major activity for research, applications, training, and education. Adding event databases (e.g., AIR France 447) to the UI to automatically populate the spatial and temporal ranges can also be very helpful.

Designing user-friendly user interfaces can be a challenge. There is no one-size-fits-all UI for users at different levels. As suggested by previous research, data repositories need to provide multiple ways (e.g., Giovanni) for data and information access. For example, NLP can provide an easy access interface for nonprofessional users or individuals addressing questions like ‘What is the average temperature in August in Paris, France?’ Furthermore, data services can be developed to provide the answer directly, other than data or tool links. In this example, it will return the average air temperature in August in Paris.

Data repositories can provide different ways to deliver data and information. In addition to NLP, Giovanni is particularly welcome in the research and education user communities. Giovanni makes it easier to use global and regional satellite and model data, as no software and data downloads are required. Likewise, more tailored tools can be developed for different communities to provide additional ways to discover and use earth data.

4.4 DATASET LANDING PAGE (DLP)

The DLP serves as a one-stop shop for data-related information and services. The DLP is still evolving. There is a need for improvements in the DLP at GES DISC to include missing or incomplete information about the datasets, such as variables, publications, user forum, FAQs, and how-to tutorial documents. Also, data metrics and user comments (or forum links with search tags) need to be added in DLP to further assist new users in using data and processing software (e.g., learning experiences from others and helping each other to answer questions about similar software and science). The current DLP is designed for an individual dataset, not a collection of multiple datasets. Likewise, there is a need for interdisciplinary data collection landing pages.

Relevant datasets and information are not linked in any DLPs at GES DISC. Using IMERG as an example, the DLP does not list the following: other similar datasets, datasets frequently used with IMERG, its related datasets (e.g., input to the algorithm), and related subject information, which is particularly relevant to facilitate heuristic search for interdisciplinary users.

4.5 CUSTOMIZED DATA

Lacking the data that users want can clearly be an issue for data discovery. For example, IMERG is a very popular precipitation data collection consisting of Early Run, Late Run and Final Run. Each run is for users with different needs to support a variety of research and applications that require different data latencies and quality. However, only two temporal resolution products are provided: half-hourly and monthly. The GES DISC recognized the need for daily IMERG products and developed three daily products that have become very popular in the user community. For those who look for hourly, three-hourly, or 10-day data, they will not find any in the GES DISC search interface or in Giovanni. Instead, they need to download either the half-hourly or daily data to develop their own hourly, three-hourly, or 10-day product.

Providing value-added customized data for users can be a part of analysis ready data (ARD) (e.g., NASA Earthdata 2022f), which can be provided not only from the cloud but also from on-premises data services. ARD can be defined as data with minimal data processing needed and the right format for immediate data analysis (e.g., visualization). Before the cloud, there were very few ARD datasets available for users due to several reasons. These reasons range from the lack of computing resources to the lack of expertise. Cloud environments enable scalable capabilities for data processing, and as a result, ARD will increasingly become available from routine data services.

4.6 METADATA

Metadata play a crucial role in providing information about datasets. Metadata comes from two sources: (1) data producers who put metadata in a dataset file and (2) DAAC staff who curate the dataset and collect them from product providers. Each dataset file often contains metadata that describes the dataset, such as the dataset producer's information, science and ancillary keywords, data quality, and variables, among others. Both sources heavily depend on data producers to provide information. During the archive process, metadata are submitted by DAAC staff based on the information received from the data producer. Meanwhile, DAAC staff can add additional information, but it is not always feasible because there are many datasets to curate. Staff members may not be the most appropriate persons to provide such information. To keep metadata complete and comply with standards, a Data Product Development Guide (DPDG) for Data Producers (NASA Earthdata 2022g) that contains a list of required and recommended metadata fields has been rolled out to collect metadata from product producers. However, the DPDG is new to the science community, and it may take time to reach maturity. Metadata standards are also important, but vocabulary varies in different disciplines and needs additional efforts to ensure usability and consistency.

4.7 ONTOLOGIES AND SEMANTIC SEARCH

Semantic websites depend on established ontologies (e.g., vocabularies, grouping of entities and their relations), which are reliant on scientific communities to develop. For example, it is well known that different vocabularies exist in different disciplines (Parsons et al. 2022), making development of vocabularies across disciplines difficult because scientists are commonly trained in and work in only one or a few related disciplines. Close collaboration between the geosciences and computational sciences to develop ontologies for interdisciplinary activities is needed. Furthermore, dedicated efforts are needed to create semantic-based methodologies, tools, and infrastructure (Narock & Fox 2015).

4.8 INFORMATION QUALITY

Information quality also plays an imperative role in data discovery. Data quality information for satellite and model-based products can be helpful (e.g., available in search results and DLPs) when many similar datasets are available. Data quality information can be obtained from instrument specifications, observation anomalies, ground validation, and more. However, obtaining such information can be a challenge, especially at a global scale for satellite and model data. For example, ground validation activities rely on in situ observations over land and ocean, which can be difficult to collect and calibrate uniformly. For interdisciplinary research, more data products are derived from multidisciplinary products, and data quality information requires additional research work to obtain such information. In addition to developing data quality information, another challenge is to provide standardized and FAIR-ready information on data quality so comparison is possible, which can take considerable effort to overcome many obstacles.

4.9 PERSONALIZED DATA SERVICES

Each user is different in terms of search habits and behavior. Developing *personalized* data services is needed to improve data and information discoverability, and this requires many efforts, such as user registration, preference storage, and software development, to enable such services. There are many benefits for both the user and the data provider from user registration. For service developers, they can better understand what users they serve (e.g., discipline, professional level) to create tailored services. For users, they can receive customized services, including dashboards (e.g., GES DISC), updates from data repositories, helpful tips from other users, and more. Nonetheless, more research is needed to create personalized data search capabilities.

4.10 APPLICABILITY OF PRACTICES TO OTHER EARTH DATA REPOSITORIES

A common concern is whether the practices for data discovery at GES DISC can be applied to other earth data repositories, such as NASA DAACS and Earthdata (NASA Earthdata 2022a). First, these practices are not designed for only a few special datasets. The data collection at GES DISC consists of satellite and model data products. Second, the CMR is used by all DAACs, including Earthdata, for NASA EOSDIS data products and services. Third, most information on a DLP or other information page is based on metadata from CMR. The DPDG provides a standard way for both data producers and service providers to generate metadata for CMR. Outside NASA, the practices can still be applicable to other data repositories because metadata play a key role in modern earth data management activities. Implementation depends on several factors, such as available resources, the level of difficulty, and user needs, among others.

4.11 EDUCATION AND TRAINING

Interdisciplinary education and training programs are important for current and future workforce development as well as awareness and outreach activities. As mentioned earlier, staff members are rarely trained to support interdisciplinary research and applications that involve the use of multiple datasets, ranging from satellite observations to model outputs. Examples, courses, and training mechanisms for the use of multiple interdisciplinary measurements need to be developed or integrated into university big data programs. Learning materials should also be made available to the user community so that users can become more knowledgeable when they search and use data. In addition, training materials need to be developed for product providers to improve awareness and enhance metadata quality, which plays a key role in data discovery in data services.

5. SUMMARY AND RECOMMENDATIONS

There are still many challenges in data discovery for interdisciplinary research, applications, and education. Over the years, efforts have been made, but progress has been limited (not as predicted or expected), particularly in the search for knowledge and personalized data services. The main challenges for practices include (1) difficulty in understanding user inquiries; (2) dataset search, other than variable search; (3) limited UI search capabilities (e.g., keywords); (4) missing or incomplete information and data and information not linked in DLP; (5) limited data products that meet users' needs; (6) missing or incomplete metadata; (7) lack of ontologies and semantic search; (8) difficulties in generating standardized information quality; and (9) lack of adequate interdisciplinary education and training programs. The scope of some of these challenges is beyond the capabilities of one data repository, requiring community efforts to be included in the long-term plan. In the short term, a repository can expand or develop services that allow users to use a hands-on, self-guided, or interactive heuristic approach to discover data and information. Based on the preceding discussion, the following recommendations are made:

1. Provide helpful suggestions in the search box.
2. Improve dataset search to variable search.
3. Enable spatial and temporal ranges and add event search to the search interface.
4. Develop a DLP for an interdisciplinary subject (e.g., air quality, wildfires).
5. Add additional helpful data information services (e.g., FAQs, user forum, data usage and limitations) to enhance heuristic search in DLP.
6. Link relevant datasets and usage information (e.g., publications, applications) in DLP.
7. Provide personalized data services.
8. Have dataset developers provide standardized metadata, including application areas and data quality information.
9. Develop metrics to measure the effectiveness of new improvements.
10. Develop interdisciplinary training and education programs for awareness and workforce development.

11. Develop ontologies and standards for interdisciplinary data services through community efforts.
12. Conduct additional research to develop a better understanding of data quality.
13. Engage a broader community in sharing and discussing best practices.

The first nine recommendations above are not limited to GES DISC and could be considered by other data repositories (e.g., Earthdata) in conjunction with their resources, priorities, and user needs. The last four recommendations require a larger scale of collaboration from scientific communities (e.g., ESIP, RDA). In short, it takes the whole community to improve data discoverability for interdisciplinary research and applications.

ACKNOWLEDGEMENTS

We acknowledge the staff at GES DISC for the development and maintenance of data services. This work and GES DISC were funded by NASA's Science Mission Directorate (SMD). We would like to thank two anonymous reviewers who provided insightful and helpful comments and suggestions.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Zhong Liu orcid.org/0000-0001-8150-7556
NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), US;
George Mason University, US

Chung-Lin Shie orcid.org/0000-0002-1115-1029
NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), US;
University of Maryland Baltimore County (retired), US

Suhung Shen
NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), US;
George Mason University, US

James Acker
NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), US;
Adnet Systems, Inc., US

Angela Li
NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), US

Jennifer C. Wei orcid.org/0000-0002-1539-2137
NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), US

David J. Meyer orcid.org/0000-0002-3090-8920
NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), US

REFERENCES

Acker, JG and **Leptoukh, G.** 2007. Online analysis enhances use of NASA earth science data. *Eos, Transactions American Geophysical Union*, 88(2): 14–17. DOI: https://doi.org/10.1029/2007EO020003

ACSI. 2022. American Customer Satisfaction Index (ACSI). Available at: https://www.theacsi.org/ [Last accessed 1 September 2022].

Augustin, H, Sudmanns, M, Tiede, D, Lang, S and **Baraldi, A.** 2019. Semantic Earth Observation Data Cubes. *Data*, 4(3): 102. DOI: https://doi.org/10.3390/data4030102

Behnke, J, Mitchell, A and **Ramapriyan, H.** 2019. NASA's Earth Observing Data and Information System—Near-term challenges. *Data Science Journal*, 18(1): 40. DOI: https://doi.org/10.5334/dsj-2019-040

Bosilovich, MG, Lucchesi, R and **Suarez, M.** 2016. *MERRA-2: File Specification*. GMAO Office Note No. 9 (Version 1.1). Greenbelt, MD: Global Modeling and Assimilation Office. Available at: http://gmao.gsfc.nasa.gov/pubs/office_notes [Last accessed 1 September 2022].

Bugbee, K, le Roux, J, Sisco, A, Kaulfus, A, Staton, P, Woods, C, Dixon, V, Lynnes, C and **Ramachandran, R.** 2021. Improving discovery and use of NASA's earth observation data through metadata quality assessments. *Data Science Journal*, 20(1): 17. DOI: https://doi.org/10.5334/dsj-2021-017

Contaxis, N, Clark, J, Dellureficio, A, Gonzales, S, Mannheimer, S, Oxley, PR, et al. 2022. Ten simple rules for improving research data discovery. *PLoS Computational Biology*, 18(2): e1009768. DOI: https://doi.org/10.1371/journal.pcbi.1009768

Devopedia. 2022. Semantic Web. Version 8, February 15. Available at: https://devopedia.org/semantic-web [Last accessed 1 September 2022].

ESIP. 2022a. Discovery Cluster. ESIP. Available at: https://wiki.esipfed.org/Discovery_Cluster [Last accessed 1 September 2022].

ESIP. 2022b. The Earth Science Information Partners (ESIP). Available at: https://www.esipfed.org/ [Last accessed 1 September 2022].

Fox, P, VSTO Team and **SeSF Team.** 2015. Semantic search in solar-terrestrial sciences. In: Narock, T and Fox, P (eds.), *The Semantic Web in Earth and Space Science: Current Status and Future Directions*. Studies on the Semantic Web, Vol. 20. Amsterdam: IOS Press. pp. 127–146 DOI: https://doi.org/10.3233/978-1-61499-501-2-127

Huffer, B, Cotnoir, M and **Gleason, J.** 2015. Ontology-drive data access at the NASA earth exchange. In: Ho, H, Chin Ooi, B, Zaki, MJ, et al. (eds.), *Proceedings: 2015 IEEE International Conference on Big Data (Big Data)*, October 29–November 1, 2015, Santa Clara, CA. n.p.: Piscataway, NJ: IEEE. pp. 2177–2181. DOI: https://doi.org/10.1109/BigData.2015.7364004

Huffman, GJ. 2022. Introduction to global precipitation algorithms and data sets. International Precipitation Working Group. Available at: http://ipwg.isac.cnr.it/data.html [Last accessed 1 September 2022].

IPWG. 2022. International Precipitation Working Group. Available at: http://ipwg.isac.cnr.it/ [Last accessed 1 September 2022].

Lafia, S, Jablonski, J, Kuhn, W, Cooley, S and **Medrano, FA.** 2016. Spatial discovery and the research library. *Transactions in GIS*, 20(3): 399–412. DOI: https://doi.org/10.1111/tgis.12235

Li, W, Goodchild, MF and **Raskin, R.** 2014. Towards geospatial semantic search: Exploiting latent semantic relations in geospatial data. *International Journal of Digital Earth*, 7(1): 17–37. DOI: https://doi.org/10.1080/17538947.2012.674561

Liu, Z and **Acker, J.** 2017. Giovanni: The bridge between data and science. *Eos*, 98. DOI: https://doi.org/10.1029/2017EO079299

Liu, Z, Shie, C-L, Ritrivi, AJ, Lei, G-D, Alcott, GT, Greene, M, Acker, J, Wei, JC, Meyer, DJ, Li, A and **Al-Jazrawi, AF.** 2022. Developing metrics for NASA earth science interdisciplinary data products and services. *Data Science Journal*, 21(1): 5. DOI: https://doi.org/10.5334/dsj-2022-005

Mathiak, B, Juty, N, Bardi, A, Colomb, J and **Kraker, P.** 2023. What are researchers' needs in data discovery? Analysis and ranking of a large-scale collection of crowdsourced use cases. *Data Science Journal*, 22(1): 3. DOI: https://doi.org/10.5334/dsj-2023-003

McGibbney, LJ, Armstrong, EM, et al. 2019. Search relevance recommendations for earth science. Technical note ESDS-RFC-037. Available at: https://www.earthdata.nasa.gov/s3fs-public/imported/ESDS-RFC-037v1.0.pdf [Last accessed 1 September 2022].

Molod, A, Takacs, L, Suarez, M and **Bacmeister, J.** 2014. Development of the GEOS-5 atmospheric general circulation model: Evolution from MERRA to MERRA-2. *Geoscientific Model Development Discussions*, 7(6): 7575–7617. DOI: https://doi.org/10.5194/gmdd-7-7575-2014

Molod, A, Takacs, L, Suarez, M, Bacmeister, J, Song, I-S and **Eichmann, A.** 2012. *The GEOS5 Atmospheric General Circulation Model: Mean Climate and Development from MERRA to Fortuna*. NASA Technical Report Series on Global Modeling and Data Assimilation, NASA/TM–2012-104606, Vol. 28.

Narock, T and **Fox, P.** (eds.) 2015. *The Semantic Web in Earth and Space Science: Current Status and Future Directions*. Studies on the Semantic Web, Vol. 20. Amsterdam: IOS Press. DOI: https://doi.org/10.3233/978-1-61499-501-2-127

NASA DAACs. 2022. EOSDIS Distributed Active Archive Centers (DAAC). Available at: https://earthdata.nasa.gov/eosdis/daacs [Last accessed 1 September 2022].

NASA Earthdata. 2022a. Earthdata—Open access for open science. Available at: https://www.earthdata.nasa.gov/ [Last accessed 1 September 2022].

NASA Earthdata. 2022b. Data Pathfinders. Available at: https://www.earthdata.nasa.gov/learn/pathfinders [Last accessed 1 September 2022].

NASA Earthdata. 2022c. Earthdata Cloud evolution. Available at: https://www.earthdata.nasa.gov/eosdis/cloud-evolution#:~:text=Further%20many%20of%20NASA's%20EOSDIS,more%20data%20being%20added%20weekly [Last accessed 1 September 2022].

NASA Earthdata. 2022d. Common Metadata Repository (CMR). Available at: https://www.earthdata.nasa.gov/eosdis/science-system-description/eosdis-components/cmr [Last accessed 1 September 2022].

NASA Earthdata. 2022e. Unified Metadata Model (UMM). Available at: https://www.earthdata.nasa.gov/unified-metadata-model-umm#:~:text=NASA's%20UMM%20is%20an%20extensible,EOSDIS%20CMR%2Dsupported%20metadata%20standards [Last accessed 1 September 2022].

NASA Earthdata. 2022f. EOSDIS data in the cloud: User requirements. Available at: https://www.earthdata.nasa.gov/learn/articles/eosdis-data-cloud-user-requirements [Last accessed 1 September 2022].

NASA Earthdata. 2022g. Data Product Development Guide for Data Producers. Available at: https://www.earthdata.nasa.gov/esdis/esco/standards-and-references/data-product-development-guide-for-data-producers [Last accessed 1 September 2022].

NASA EOSDIS. 2022a. Earth Observing System Data and Information System (EOSDIS). Available at: https://earthdata.nasa.gov/eosdis [Last accessed 1 September 2022].

NASA EOSDIS. 2022b. American Customer Satisfaction Index (ACSI) reports. Available at: https://earthdata.nasa.gov/eosdis/system-performance/acsi-reports [Last accessed 1 September 2022].

NASA GES DISC. 2022a. NASA Goddard Earth Sciences Data and Information Services Center (GES DISC). Available at: https://disc.gsfc.nasa.gov [Last accessed 1 September 2022].

NASA GES DISC. 2022b. Migrating to the cloud. Available at: https://disc.gsfc.nasa.gov/information/documents?title=Migrating%20to%20the%20Cloud [Last accessed 1 September 2022].

NASA GES DISC. 2022c. How to obtain data for conducting hurricane case study. Available at: https://disc.gsfc.nasa.gov/information/howto?keywords=hurricane&title=How%20to%20Obtain%20Data%20for%20Conducting%20Hurricane%20Case%20Study [Last accessed 1 September 2022].

NASA Giovanni. 2022. NASA Giovanni. Available at: https://giovanni.gsfc.nasa.gov [Last accessed 1 September 2022].

Parsons, MA, Katz, DS, Langseth, M, Ramapriyan, H and **Ramdeen, S.** 2022. Credit where credit is due. *Eos*, 103. DOI: https://doi.org/10.1029/2022EO220239

Ramapriyan, H and **Behnke, J.** 2020. NASA's Earth Observing System Data and Information System (EOSDIS) and FAIR—A self-assessment. In: *IN044—Improving Infrastructure for Trustworthy Digital Repositories to Enable Current and Future Use of Open Data in Developed and Developing Countries I. AGU Fall Meeting*, December 1–17, 2020.

Raskin, RG and **Pan, MJ.** 2005. Knowledge representation in the semantic web for earth and environmental terminology (SWEET). *Computers & Geosciences*, 31(9): 1119–1125. DOI: https://doi.org/10.1016/j.cageo.2004.12.004

RDA. 2022a. The RDA Data Discovery Paradigms Interest Group. Available at: https://www.rd-alliance.org/groups/data-discovery-paradigms-ig [Last accessed 1 September 2022].

RDA. 2022b. The Research Data Alliance (RDA). Available at: https://www.rd-alliance.org/about-rda [Last accessed 1 September 2022].

Stoyanova, K, Gerasimov, I, Mehrabian, A, Jahoda, E, Wei, J, Pham L and **Khayat, MG.** 2021. Application of a dataset-publication knowledge graph for improving earth science data search. In: *IN45E—Best Practices and Realities of Research Data Repositories III Poster. AGU Fall Meeting*, New Orleans, LA, December 13–17, 2021.

Wang, C, Ma, X and **Chen, J.** 2018. Ontology-driven data integration and visualization for exploring regional geologic time and paleontological information. *Computers & Geosciences*, 115: 12–19. DOI: https://doi.org/10.1016/j.cageo.2018.03.004

Wang, S, Wang, J, Zhan, Q, Zhang, L, Yao, X and **Li, G.** 2023. A unified representation method for interdisciplinary spatial earth data. *Big Earth Data* 7(1): 136–155. DOI: https://doi.org/10.1080/20964471.2022.2091310

Weikum, G. 2013. Data discovery. *Data Science Journal*, 12: pp. GRDI26–GRDI31. DOI: https://doi.org/10.2481/dsj.GRDI-005

Wikipedia. 2022. Air France Flight 447. Available at: https://en.wikipedia.org/wiki/Air_France_Flight_447 [Last accessed 1 September 2022].

Wilkinson, M, Dumontier, M, Aalbersberg, I, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Wu, M, Psomopoulos, F, Khalsa, SJ and **de Waard, A.** 2019. Data discovery paradigms: User requirements and recommendations for data repositories. *Data Science Journal*, 18(1): 3. DOI: https://doi.org/10.5334/dsj-2019-003

Wu, W-S, Purser, RJ and **Parrish, DF.** 2002. Three-dimensional variational analysis with spatially inhomogeneous covariances. *Monthly Weather Review*, 130: 2905–2916. DOI: https://doi.org/10.1175/1520-0493(2002)130<2905:TDVAWS>2.0.CO;2

TO CITE THIS ARTICLE:
Liu, Z, Shie, C-L, Shen, S, Acker, J, Li, A, Wei, JC and Meyer, DJ. 2023. Improving NASA's Earth Satellite and Model Data Discoverability for Interdisciplinary Research, Applications, and Education. *Data Science Journal*, 22: 9, pp. 1–16. DOI: https://doi.org/10.5334/dsj-2023-009

Submitted: 17 November 2022
Accepted: 28 February 2023
Published: 28 April 2023

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.