Eos

Creating Data Tool Kits That Everyone Can Use

Earth scientists outline challenges to making the growing wealth of available data more accessible and to using data services for interdisciplinary research and applications.

By **Z. Liu**, **V. Mantas**, **J. Wei**, **M. Jin**, and **D. Meyer**

11 May 2020

Multiple special or discipline-oriented tools, often with steep learning curves, are required to handle heterogeneous, complex, and evolving Earth science data sets in interdisciplinary research and applications. Credit: Pixabay/Deborah Breen Whiting

As Earth science and the technologies it uses evolve and improve, the data and services that support the science also change and become more complex, often spanning multiple disciplines. The ability to easily find and seamlessly access these data and services in an open and integrated environment is essential to facilitating interdisciplinary research and applications and to broadening data user communities.

The sheer amount of available data is growing rapidly as the science community adopts the Findable, Accessible, Interoperable, and Reusable (FAIR) data principles [***Wilkinson et al.***, 2016] and emerging technologies such as cloud computing. Even with recent advances in data archiving and services (e.g., more data sets and related information are available online with customized data services and multiple data access methods), accessing **heterogeneous** interdisciplinary data sets (e.g., those with nonuniform data types and formats) still poses challenges to users globally.

To address these challenges in Earth science interdisciplinary data services, we organized and led sessions entitled "Data and Information Services for Interdisciplinary Research and Applications in Earth Science" at AGU Fall Meetings 2018 and 2019; at both meetings, this was one of the largest of the Earth and Space Science Informatics sessions. International groups of participants presented data, tools, and services for Earth science interdisciplinary research and applications, as well as work on other topics related to big data, artificial intelligence (AI), machine learning (ML), and natural language processing.

Six Challenges

As a result of the presentations, discussions, and feedback from our AGU Fall Meeting sessions, we identified the following questions that address challenges in making Earth science data and data services more accessible and useful:

- How can we make Earth science interdisciplinary data sets needed for a specific research project or application easier to find?

- How do we eliminate the need for the many special tools, some with steep learning curves, that are currently required to handle heterogeneous and interdisciplinary data sets?
- What data services can we provide in the cloud environment, where unprecedented access to data sets and data analytics are available?
- What data services can we provide to facilitate AI/ML activities?
- How can we collect metrics to help development and enhancement of data services and to benchmark the performance and societal impacts of a project or mission?
- How do we ensure the scientific reproducibility of Earth sciences research?

Finding and Accessing the Data

Data users typically consult various sources (e.g., the Internet, conference proceedings, colleagues, professors) to find where data are archived and distributed among many repositories around the world. However, access websites for different data resources are often designed differently, and as a result, only those who are already familiar with the repositories can easily locate data sets and information. Finding the right data and information for a specific research project or application is another challenge, especially for inexperienced data users who may not be familiar with data sets outside their own disciplines and for people searching across disciplines.

> *Access websites for different data resources are often designed differently, and as a result, only those who are already familiar with the repositories can easily locate data sets and information.*

The variety of access website designs can impede data and information searches by users who are unfamiliar with specific repositories. For example, NASA's Earth Observing System Data and Information System (EOSDIS) has 12 discipline-oriented Distributed Active Archive Centers (DAACs) that archive and distribute NASA Earth data sets from satellite missions and

projects, each with its own unique web interface. For users doing interdisciplinary research that requires data from multiple DAACs, it can be difficult to become familiar with all of the interfaces.

One solution is to develop an integrated and uniform web interface for data access. At present, NASA EOSDIS is developing such an interface, called Earthdata, which serves as a gateway for all data sets and services at the 12 NASA DAACs. When the interface is finished, users will be able to search all NASA Earth science data, along with data services and information, in one place. Building this type of one-stop shop for accessing complex and heterogeneous data, services, and information is a major challenge in improving interdisciplinary research and applications.

> *Incorporating human expertise with artificial intelligence and machine learning technologies such as natural language processing may improve the user experience of finding data for a specific need.*

Another barrier to easy access is that discipline-oriented websites currently exist for their own special disciplinary requirements. To unify these different websites, an integrated data system must address both general and discipline-specific requirements. On a larger scale, Earth science data from various U.S. federal agencies, countries other than the United States, private companies, and citizen scientists must also be easily accessible in an integrated environment. Ideally, all these data would be accessible without the need to visit different websites, but making this a reality requires collaboration from domestic and international data scientists, developers, and stakeholders to address such issues as disciplinary vocabulary, data standards, and usability.

At present, many websites rely on sorting and filtering of search results. In satellite data services, for example, search suggestions, research subject, measurement, satellite source, and processing level are often used to narrow the list of search results. A user's success in finding the data they need can vary significantly in such systems, depending on the web design, the user's knowledge, and many other factors. Websites that focus on a single project or mission and contain only a few data sets can eliminate the need for sorting or filtering. For more comprehensive resources, inexperienced users

often need human assistants or a help desk to interact with them, find out more about what they need, and provide recommendations for data products or services. Therefore, incorporating human expertise with AI/ML technologies such as natural language processing in the system may improve the user experience of finding data for a specific need.

Simplifying the Tool Kit

Because interdisciplinary data sets are complex, and their formatting and data structures are not uniform, multiple tools are needed to handle such data sets for research and applications. For example, more than 61 **data tools** are available at the 12 discipline-oriented NASA DAACs for search and order; data handling; subsetting and filtering; geolocation, reprojection, and mapping; and data visualization and analysis. It can be a daunting task for a user to learn all of these tools for interdisciplinary activities.

Heterogeneous data can also present challenges to users or stakeholders without access to complex data set processing capabilities. For many users, acquiring dedicated software, using multiple tools, and performing programming-based data analyses are not viable. The use of standards like uniform data formatting may address the problem of heterogeneous data often requiring many tools to handle. Data tools that integrate more data processing capabilities may also help in reducing the number of tools. Meanwhile, data repositories can go beyond their existing services (e.g., providing original data as they are) by offering data interoperability services to provide data that meet users' research or application requirements with respect to data format, projection, model grid, and spatiotemporal resolution, for example.

Putting New Capabilities to Work

Cloud computing provides new opportunities to address issues related to the unprecedentedly large amount of **data and data analytics** available. Governmental and private organizations are putting significant efforts toward developing cloud-based data services. For example, the NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) plans to launch its popular online **visualization and analysis tool** Giovanni in the cloud, creating the potential to scale up and expand its current capabilities by, for example, including Earth science data sets from other NASA DAACs, improving performance, and providing new data analytics.

Best practices, including user-friendly features and services, should be carried over to the cloud environment. These practices can facilitate a smooth or seamless transition from on-premises data services to the cloud and ensure a satisfactory user experience in the cloud-based environment. One such initiative currently under development is the NASA **EOSDIS Cumulus Project**, a cloud-based framework for NASA EOSDIS data collections. This project is designed so that users will not notice any difference between the on-premises and cloud-based data services.

Best practices, including user-friendly features and services, should be carried over to the cloud environment.

Cloud computing isn't the only area in which new technologies are introducing significant changes. In recent years, the Earth science community, like in many other sectors and scientific fields, has experienced a surge in research and applications using AI/ML techniques. Identifying and adopting the features that data repositories can provide to facilitate AI/ML activities are pressing and challenging issues. For example, natural language processing-based data systems could simplify access for users who want only visualizations (e.g., images, maps), facts, or information. These systems handle the interactions between computers and humans using natural language—ordinary human speech in the form of voice or text rather than arcane computer commands. On the other hand, more advanced users expect data repositories to provide analysis-ready or customized data (e.g., training data) for AI/ML activities. Down the road, standard or customized AI/ML services—running AI algorithms such as deep learning or random forests, for example—can be integrated into data repositories, allowing users to conduct AI/ML activities without leaving the system. Cloud computing may be able to host such services in the future.

Measuring Performance and Impact

Data metrics are frequently used to measure user and system activities—data access and usage in research and applications, for example—that are related to the life cycle of a data set and play an important role in Earth science. For example, in the satellite community, data usage metrics are used to benchmark a mission's or project's success and its societal impacts.

Among their many applications, data metrics supply key information to satellite data service providers for designing new data services and improving existing ones. Satellite product developers rely on data metrics to understand how their products are used. Thus, collecting data metrics is an essential part of a satellite mission or project.

> ***One key challenge is to develop metric standards for different disciplines so that metrics are interoperable in interdisciplinary activities.***

One key challenge is to develop metrics that accurately describe a wide range of data-related activities in research and applications. Another is to develop metric standards for different disciplines so that metrics are interoperable in interdisciplinary activities.

But Is It Reproducible?

Scientific reproducibility is a cornerstone of Earth science (and of all scientific fields). Providing trustworthy data and results is one of the most challenging and pressing issues in the science community. Reproducing Earth science research requires documentation of all elements in the life cycle, including data, algorithms, software, computing environment, and other factors.

Several research projects and workshops have focused on ways to improve reproducibility by, for example, collecting best practice guidelines—like the community guidelines for open and reproducible workflows that the geoscience research modeling community has worked to develop [*Mullendore et al.*, 2020]. Earth science users are also increasingly using open-source software, open standards, and services such as Jupyter Notebook to exchange workflows. For interdisciplinary research and applications, the scope is much larger, requiring participation and collaboration from the entire Earth science community and from stakeholders (e.g., journal publishers). Major challenges include the development of standards for the framework and for interoperability.

Going Far, Together

There are many challenges to improving Earth interdisciplinary data services. Emerging technologies like cloud computing and AI/ML can potentially enable significant progress in all aspects of interdisciplinary data services. However, tackling the challenges will also depend on active participation and collaboration from all involved parties in the community. In the wake of AGU Fall Meeting 2019, we plan to continue our follow-up activities, which include a special journal issue to report current advances and address the challenges.

Acknowledgments

We thank all session participants of the past two AGU Fall Meetings. GES DISC is funded by NASA's Science Mission Directorate. Vasco Mantas received funding from the European Union's Horizon 2020 program under grant agreement GA 776026.

References

Mullendore, G. L., M. S. Mayernik, and D. Schuster (2020), Determining best practices for archiving and reproducibility of model data, paper presented at the 36th Conference on Environmental Information Processing Technologies, Am. Meteorol. Soc., Boston, Mass., 12–16 Jan.

Wilkinson, M. D., et al. (2016), The FAIR guiding principles for scientific data management and stewardship, *Sci. Data, 3*, 160018, https://doi.org/10.1038/sdata.2016.18.

Author Information

Zhong Liu (zhong.liu@nasa.gov), NASA Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, Md.; also at Center for Spatial Information Science and Systems, George Mason University, Fairfax, Va.; Vasco Mantas, University of Coimbra, Coimbra, Portugal; Jennifer Wei, NASA GES DISC, Greenbelt, Md.; Menglin Jin, University of

Maryland, College Park; and David Meyer, NASA GES DISC, Greenbelt, Md.

Citation:

Liu, Z., V. Mantas, J. Wei, M. Jin, and D. Meyer (2020), Creating data tool kits that everyone can use, *Eos, 101*, https://doi.org/10.1029/2020EO143953. Published on 11 May 2020.

Text © 2020. The authors. CC BY-NC-ND 3.0
Except where otherwise noted, images are subject to copyright. Any reuse without express permission from the copyright owner is prohibited.

© 2022 American Geophysical Union. All rights reserved.