feature article

Open-Source Science: The NASA Earth Science Perspective

Kevin Murphy, NASA Headquarters, kevin.j.murphy@nasa.gov

Since its establishment, NASA has acquired and provided data about space and Earth's atmosphere to foster scientific research. These data are integral components of research into Earth's interconnected systems, and NASA Earth science data have been openly available to all users since the Earth Observing System Data and Information System (EOSDIS) became operational in 1994 as a key component of NASA's Earth Observing System (EOS).[1] Further, since 2015 data systems software developed through NASA research and technology grants and awards has been made available as *open-source software*,[2] which means that the source code for these tools is freely available for inspection, modification, and enhancement.[3] These policies and practices enable anyone anywhere in the world to access more than 57 petabytes (PB) of NASA Earth science data—one of the largest repositories of Earth science data on the planet—fully, openly, and without restriction.[4]

The development of open-source software fundamentally changed how software was shared, and enabled software and code to be available more broadly and shared collaboratively with diverse groups to accelerate software development. These features of the open-source software movement are key attributes of what is known as *open science*, which is defined as "a collaborative culture enabled by technology that empowers the open sharing of data, information, and knowledge within the scientific community and the wider public to accelerate scientific research and understanding."[5]

A system based on open science aims to make the scientific process as transparent (or *open*) as possible by making all elements of a claimed discovery readily accessible, which enables results to be repeated and validated. Out of this open science concept, an evolving scientific paradigm called *open-source science* is emerging.

Open-source science builds on concepts from the open-source software revolution that expanded participation in code development and applies these concepts to the scientific process to accelerate discovery by conducting science openly from project initiation through implementation. The result is the inclusion of a wider, more diverse community in the scientific process as close to the start of research activities as possible. This increased level of commitment to conducting the full research process openly and without restriction enhances transparency and reproducibility, which engenders trust in the scientific process. It also represents a cultural shift that encourages collaboration and participation among practitioners of diverse backgrounds, including scientific discipline, gender, ethnicity, and expertise. Open-source science is more equitable science.

A system based on open science aims to make the scientific process as transparent (or open*) as possible by making all elements of a claimed discovery readily accessible, which enables results to be repeated and validated.*

[1] The story of the evolution of EOSDIS (up to 2009) was told in a two-part article, "EOS Data and Information System, Where We Were and Where We Are," that appeared in the July–August 2009 and September–October 2009 issues of *The Earth Observer* [**Volume 21, Issue 4**, pp. 4–10 and **Volume 21**, **Issue 5**, pp. 8–15—*go.nasa.gov/3uzO6AL*].

[2] For a detailed description of NASA's Earth Science Data Systems (ESDS) Program open-source software policy, see *go.nasa.gov/2WkvlEW*.

[3] For a broad discussion of NASA Earth science data operations, including the EOSDIS Distributed Active Archive Centers (DAACs), see "Earth Science Data Operations: Acquiring, Distributing, and Delivering NASA Data for the Benefit of Society" in the March–April 2017 issue of *The Earth Observer* [**Volume 29**, **Issue 2**, pp. 4–18—*go.nasa.gov/3kKFOTy*]. For an overview of the DAACs and a review of their milestones, see *go.nasa.gov/3uhu5yN*.

[4] At the end of August 2021, the total EOSDIS archive volume was 57.2 PB. To learn more, visit *go.nasa.gov/3ueDGGL*.

[5] This definition and more information on other topics discussed in this article can be found in a 2021 article by Rahul Ramachandran, Kaylin Bugbee, and Kevin Murphy: "From Open Data to Open Science," *Earth and Space Science* [**Volume 8**, **Issue 5**—*doi:10.1029/2020EA001562*].

feature article

The SMD and ESDS vision is to use open-source science principles to expand participation in the scientific process, improve reproducibility, and accelerate scientific discovery.

Open-Source Science in NASA's Science Mission Directorate

Open-source science is a foundational objective of NASA's Science Mission Directorate (SMD) and SMD's Earth Science Data Systems (ESDS) Program.[6] Along with the wide dissemination and use of openly available Earth-observing data, the SMD promotes and facilitates the full and open sharing of all *metadata* (information that describes data), documentation, models, images, and research results achieved using these data and makes available the source code used to generate, manipulate, and analyze the data.[7] The SMD and ESDS vision is to use open-source science principles (described below) to expand participation in the scientific process, improve reproducibility, and accelerate scientific discovery.

Three primary elements work together to fulfill open-source science objectives:

- Open access to data, software, and any information coming out of research, such as journal articles, blog posts, and similar products as early in the scientific process as possible;
- open access to the scientific process with transparency and active inclusion of different communities; and
- a collaborative and inclusive process that is welcoming and open to everyone.

The definition of open-source science and its foundational elements imply four distinct meanings to *open*. Open means *transparent*, in that scientific results and processes should be visible, accessible, and reproducible to a wide audience. Open also means *inclusive* and welcoming to participation by and collaboration with a diverse range of people and organizations. Further, open implies *accessible* to all users with *reproducible* processes and results.

Open-Source Science as an Evolving Paradigm

The development of open-source science has been aided by the growth of *big data* collections,[8] cloud-based computing systems, and open-source web applications and environments; increases in personal-computer processing power; and the ability to work close to cloud-based data collections with little more than a reliable internet connection.[9]

As noted earlier, the shift to open-source science is a change from more traditional science systems that have barriers to participation in the scientific process—see **Figure 1** on next page.

These barriers include:

- Restrictions on algorithm or code sharing;
- requirements for expensive processing equipment or systems;

[6] For more detail, please see "SMD's Strategy for Data Management and Computing for Groundbreaking Science: 2019 to 2024," which can be downloaded from *go.nasa.gov/3APYuGX*.
[7] To learn more, see "SMD Policy Document (SPD)-41: Scientific Information Policy for the Science Mission Directorate," which can be downloaded from *go.nasa.gov/2ZCCBgE*
[8] *Big data* refer not only to large-volume data collections, but also to methods developed to extract information from datasets that are too large to be processed using traditional data-processing software.
[9] The ability to work "close" to or "next" to big data collections in a cloud-based data system enables anyone with an internet connection to conduct their analyses and work directly with data in the cloud without having to download or store data. After working with the cloud-based data collection, a researcher needs only to download the results of their analysis—a significant savings in time, cost, and computing power. Having data collections in the cloud also facilitates collaborative work on the same data collection simultaneously by multiple research teams in different locations.

- retention of data for exclusive use by science teams, or delays in sharing research-quality data;
- restrictive journal publishing and research dissemination processes; and
- policies and employment, communication, and collaboration strategies that favor particular groups through stated or unstated preferences for educational level, gender, professional affiliation, geographic location, and other personal attributes.

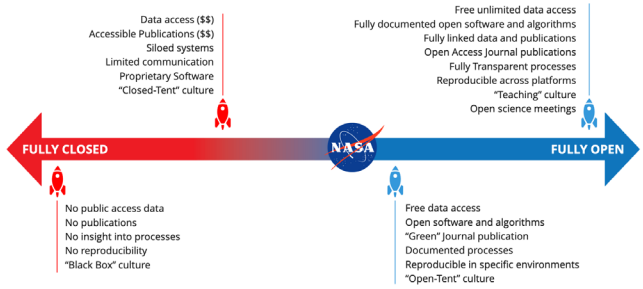

Figure 1. NASA's provision of all mission data, code, algorithm theoretical basis documents (ATBDs), and analysis software places the agency (indicated by the NASA logo) closer to the *Fully Open* side of the continuum of closed [*left*] vs. open [*right*] science systems. Continued enhancements to data systems and architecture—along with greater acceptance of open-source science principles—will help the agency move closer to the Fully Open ideal. **Image credit:** NASA ESDS

NASA SMD and ESDS data policies and practices require that all NASA-funded researchers use open-source software for any code developed as part of the research process. To make this as easy as possible for science teams, the agency provides guidance for the appropriate licensing that needs to be applied to code to ensure it is fully open. In addition, ESDS requires that all developed code be delivered to a publicly accessible repository service that is widely recognized by a large, active, open-source software community and used by developers of Earth science data and tools. ESDS encourages NASA-funded researchers to deliver code and supporting algorithms to the NASA GitHub Repository at *github.com/nasa*.

ESDS also provides resources for standardizing mission code and software. An example of one effort to accomplish this is the development of the Algorithm Publication Tool (APT) for algorithm theoretical basis documents (ATBDs).[10] Created by NASA's Interagency Implementation and Advanced Concepts Team (IMPACT), a prototype of the APT was developed in 2019; the system is now in its second phase of development.[11] When fully operational, the APT will be an important step toward enabling open, reproducible science by helping scientists write standardized, high-quality algorithm documentation collaboratively and provide functionality to make ATBDs open-access literature. A primary goal of the APT is to provide a free and open portal to ensure that all ATBDs are discoverable and accessible to users. High-quality supporting metadata, populated during the ATBD publication phase, will allow users to easily search for documents and the content within so that the most relevant information is readily discoverable.

As noted in NASA's Data and Information Policy, the agency is committed to the full and open sharing of Earth science data obtained from NASA's Earth-observing satellites, suborbital platforms, and field campaigns with all users as soon as these data become available.

Open-Source Science in Practice

An integral element of open-source science is the public provision of data and code as early in the scientific or mission development process as possible. As noted in NASA's Data and Information Policy, the agency is committed to the full and open sharing of Earth science data obtained from NASA's Earth-observing satellites, suborbital platforms, and field campaigns with all users as soon as these data become available. In addition, there is no period of exclusive access to NASA Earth science data. Following a postlaunch checkout period, all data are made available to public user communities.[12]

[10] ATBDs exist for each data product and describe the procedures used to create them. Examples can be found at *go.nasa.gov/3m6utMK*.
[11] More information about the APT, including background papers and the current state of the system's development, is available through the IMPACT website at *go.nasa.gov/2Y4C5Yb*.
[12] To learn more see, the NASA SMD Scientific Information Policy (SPD-41), which is referenced in footnote 7.

feature article

Open-source science provides numerous benefits to scientists and those involved in the scientific process. By removing barriers to participation, the diversity of those engaged in the process will increase.

Finally, the publication of results acquired from the use of these data is accomplished using open-access *preprint servers*—online repositories established specifically to disseminate scientific research results or data associated with scholarly papers that are not yet peer reviewed or accepted by traditional academic journals. Preprint servers enable researchers to share initial scientific results with colleagues and receive feedback while a full article is undergoing prepublication peer review. An extensive list of preprint servers organized by discipline, external content indexing, permanence and preservation of content, and other criteria is available through the nonprofit Accelerating Science and Publication in biology (ASAPbio) at *asapbio.org/preprint-servers*.

Benefits of Open-Source Science

Through the implementation of the features previously noted, open-source science provides numerous benefits to scientists and those involved in the scientific process. By removing barriers to participation, the diversity of those engaged in the process will increase. Further, making all data, code, algorithms, and supporting documentation openly available in a variety of locations (through, for example, NASA Earthdata Search,[13] EOSDIS Distributed Active Archive Centers (DAACs), NASA Open Innovation Sites,[14] and the NASA Technical Reports Server[15]) makes the scientific process more transparent and scientific results more reproducible. This, in turn, lends further legitimacy to the scientific process. In addition, the use of standardized mission code and software makes it easier for new users to learn to interact with the data and helps foster greater participation in the scientific process. Finally, broadening the base of those able to work with, analyze, and use Earth science data enables research results to be applied more quickly and broadly for societal benefit.

NASA Open-Source Science in Practice: MAAP

One example of an ESDS-sponsored open-source science effort is the joint NASA/European Space Agency (ESA) Multi-Mission Algorithm and Analysis Platform (MAAP)—see **Figure 2**. MAAP integrates biomass data from multiple missions operated by different space agencies into a consistent, cloud-based data record that can be openly used by global stakeholders.[16] MAAP code is written in *Jupyter Notebooks* [an

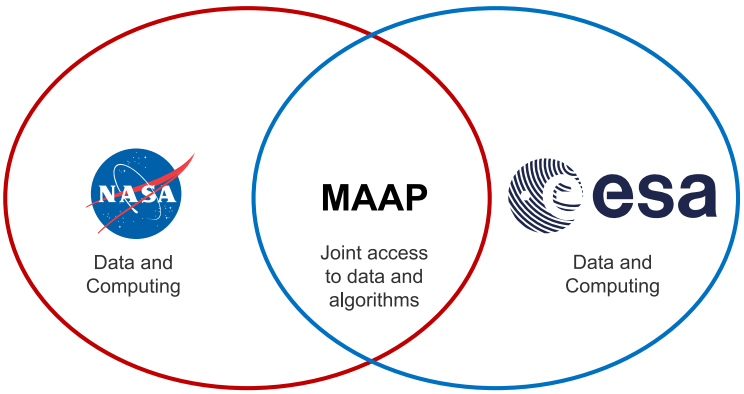

Figure 2. MAAP is a virtual, open, and collaborative environment that leverages cloud technologies to facilitate open data use across aggregated datasets. Using MAAP, NASA and ESA are working together to make terrestrial biomass data and metadata from multiple missions and sources more interoperable across organizations. To learn more, visit *earthdata.nasa.gov/esds/maap*. **Image credit:** NASA MAAP

[13] To learn more, see *go.nasa.gov/3ic97wu*.
[14] NASA has Open Innovation Sites for data (*data.nasa.gov*), code (*code.nasa.gov*), and APIs (*api.nasa.gov*).
[15] The server can be found at *ntrs.nasa.gov*.
[16] Data that will be part of MAAP include the current NASA Global Ecosystem Dynamics Investigation (GEDI) mission and the joint NASA/ESA AfriSAR airborne campaign as well as the upcoming ESA Biomass and joint NASA/Indian Space Research Organisation Synthetic Aperture Radar (NISAR) missions.

open document format based on JavaScript Object Notation (JSON)] that are openly shared between teams using the MAAP platform. Having all MAAP data, code, and infrastructure openly available speeds up the research process and facilitates efficient collaboration. As **Laura Duncanson** [University of Maryland, College Park—*MAAP Project Scientist*] observes, "Now all of us can learn from each other's code. The platform feels like a true paradigm shift; it's the right way forward."

A public announcement of MAAP Version 1 is scheduled for fall 2021, with MAAP Version 2 scheduled for release in spring 2022.

TOPS Steers NASA Toward Open Science

NASA is forming a steering team for an upcoming open-source science initiative: the Transform to Open Science (TOPS). Scheduled to kick off in October 2021, TOPS will coordinate efforts designed to rapidly transform the way NASA researchers—and researchers at other agencies, organizations, and communities—do their work. These efforts will be aligned with SMD's *Strategy for Data Management and Computing for Groundbreaking Science 2019-2024*[17] and further enabled by the National Academies of Sciences, Engineering, and Medicine (NASEM) and United Nations Educational, Scientific and Cultural Organization (UNESCO)[18] recommendations, which are intended to inform NASA's pathway forward to advance open science.

As part of TOPS, 2023 will be designated the Year Of Open Science (YOOS)—a global community initiative to spark change and inspire engagement in open science through events and activities that will help further develop the open-source science paradigm.

As part of TOPS, 2023 will be designated the Year Of Open Science (YOOS)—a global community initiative to spark change and inspire engagement in open science through events and activities that will help further develop the open-source science paradigm.

Open-Source Science in NASA's Earth System Observatory

Open-source science will be a key attribute of NASA's Earth System Observatory (ESO).[19] This array of Earth-observing missions will provide vital information to guide decisions related to climate change, severe weather and other natural hazards, wildfires, and global food production, all in keeping with the recommendations of the 2017 Earth Science Decadal Survey produced by NASEM.[20] NASA missions will be augmented with competitively selected Earth Explorer missions that will bring further innovation and additional key observations to the ESO.

ESO missions will generate greater volumes of data than any previous NASA missions. As stated earlier, the NASA EOSDIS archive volume at the end of August 2021 was more than 57 PB. The first ESO mission alone—the joint NASA/Indian Space Research Organisation Synthetic Aperture Radar (NISAR) mission (scheduled for launch in 2023)—is expected to generate more than 30 PB of data *per year*. NISAR data will help address a variety of complex environmental processes, including ice-sheet collapse and natural hazards such as earthquakes, volcanoes, and landslides.

As part of our commitment to open-source science, NASA will make all ESO mission data, code, and supporting documents available as early in the mission life cycle as feasible. Given the high volume of ESO data, these data will be stored using cloud-based systems and tools will be provided for working with these data directly in the cloud. This strategy will expand the ability of global research teams to collaboratively work with and conduct research using more NASA Earth science data than ever before. The result will be the availability of these data to a broader, more diverse global community of users with the attendant increase in opportunities for scientific discovery.

[17] A link to this document is available in Footnote 6.

[18] The UNESCO Recommendation on Open Science can be found at *en.unesco.org/science-sustainable-future/open-science/recommendation.*

[19] To learn more about the ESO, see *go.nasa.gov/3wmt4pm*.

[20] For details, see the 2017 document "Thriving on Our Changing Planet: A Decadal Strategy for Earth Observations from Space," available for download from *go.nasa.gov/2wXJn2n*.

Conclusion

Open-source science is the foundation of SMD and ESDS efforts to expand the use of NASA Earth science data to a more diverse, inclusive base of users. This evolving paradigm represents not only a new way of doing science, but a new way of thinking about what science means in terms of who can participate in the scientific process. Providing mission data, code, and supporting documents fully and openly—and as early in the scientific process as possible—broadens potential participation, enables collaborative work with big

Acknowledgements

The author of this article, **Kevin Murphy**, is the Chief Science Data Officer for NASA's Science Mission Directorate (SMD) and the Program Manager for NASA's Earth Science Data Systems (ESDS) Program. Murphy would like to thank **Josh Blumenfeld** [NASA's Goddard Space Flight Center, ESDS Communications Team—*Managing Editor*] and the ESDS Communications Team for their contributions to this article. ■