From Open Data to Open Science

Rahul Ramachandran[1], Kaylin Bugbee[1], and Kevin Murphy[2]

[1]NASA/MSFC

[2]NASA/HQ

Corresponding author: Rahul Ramachandran (rahul.ramachandran@nasa.gov)

Key Points:

- This paper defines open science 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
- This paper provides a synopsis of various open science activities happening throughout the community and synthesizes those activities around three broad open science focus areas
- The paper also describes how data programs are in a unique position to set strategic directions that support and promote open science within the three focus areas

Abstract

The open science movement continues to gain momentum, attention and discussion. However, there are a number of different interpretations, viewpoints and perspectives as to what the term 'open science' means. In this paper, we define open science 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. Opportunities exist for data programs to encourage and accelerate open science through thoughtful design. This paper provides a synopsis of various open science activities happening throughout the community and synthesizes those activities around three broad open science focus areas. The paper then describes how data programs are in a unique position to set strategic directions that support and promote open science within the three focus areas. The paper also discusses the challenges the community faces in promoting open science. The paper closes with a call to action for both data programs and individual researchers to support the wider adoption of open science principles.

1. Introduction

1.1. Open Science Definition

Open science, as both a concept and a term, is increasing in popularity and usage. However, definitions, interpretations and perceptions as to what the term 'open science' means varies. Some definitions are fairly narrow and only focus on providing more open access to science as a body of knowledge. These narrow definitions place an emphasis on openly sharing scientific knowledge as early as possible in the research process (University of Cambridge, n.d.). On the other hand, broader definitions of open science acknowledge that science is both a body of knowledge and a systematic method for thinking. Broad definitions place an emphasis on encouraging a culture of openness (Bartling & Friesike, 2014) that includes the entire process of conducting science (National Academies of Sciences, Engineering, and Medicine, 2018) and encourages open collaboration and access to knowledge (Vicente-Saez & Martinez-Fuentes, 2018). In its broadest definition, the term 'open science' refers to a paradigm shift in how the methods of science are conducted. This expansive vision of open science acknowledges that rapid technology changes, primarily driven by the Internet, may enable a second scientific revolution that fundamentally changes research methods and standards across science.

To complicate matters, the term "open science" is sometimes used interchangeably to represent various principles that support the broader idea of open science itself. These principles include ideas such as open data, open source software, open journal access and reproducibility. For example, reproducibility, or the ability to verify another scientist's results, is enabled by the open principles of open data, open code and transparent methodologies, yet reproducibility itself is not equivalent to open science.

While open science definitions are variable and ambiguous, the value of open science as both a concept and a paradigm change is accepted by the majority of the scientific community. Open science not only benefits the scientific endeavor itself but has also been shown to benefit individual researchers through increased citations and media attention, a larger collaborative network, and exposure to new career and funding opportunities (McKiernan et al., 2016; Murphy et al., 2020). However, in order for researchers, institutions and programs to more effectively foster and enable open science, a workable definition of open science is needed. This need is especially important for data programs, which are in a unique position to create policies, processes and systems that empower and promote open science. Therefore, we define open science 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. This vision of open science focuses on three overarching dimensions: (1) increasing the accessibility to the scientific process and the corresponding body of knowledge; (2) making both the research process and knowledge sharing more efficient; and (3) understanding and assessing scientific impact through innovative new metrics.

The goals of this paper are twofold. First, the paper provides an overview of the current state of various open science activities across the community. This overview is not meant to be comprehensive in nature but is instead meant to highlight the primary open science focus area at this time. Second, since the majority of science is now data driven, this paper specifically highlights how the path to open science for the Earth system science community is deeply intertwined with data programs. Data programs at different agencies and organizations play a pivotal role in accelerating the path to open science. New technology innovations for data systems and tools, along with data stewardship practices and policies, are key to making the open science paradigm shift a reality. This paper is not meant to prescribe specific data program solutions but instead hopes to stimulate discussion around new approaches along with novel ways to address some of the known issues surrounding open science.

2. Drivers for More Open Science

While science has, in some measure, been open since the 1700s, three key drivers are accelerating the open science movement and redefining what it means to be open. The three factors driving this movement are technology advancements, the rapid growth in data volume and variety (Xu & Yang, 2014) and the increasing complexity of the science questions being tackled (Hansen, 2012). These three factors are not independent but rather are profoundly intertwined, with data and technology being especially synergistic.

Rapid technology advancements are fundamentally changing how science is conducted and in tandem are accelerating the adoption of open science principles. Technology has enabled new workflows that not only make the process of science more efficient but also create new mediums for sharing knowledge with other scientists and the broader community . Furthermore, technological innovations are continually evolving the scientific process itself. New collaboration technologies make sharing ideas, data, algorithms, software and experiments easier (Friesike et al., 2015), while new software tools are now available that quickly incorporate the latest improvements in algorithms and analysis methods such as machine learning libraries (Woelfle et al., 2011). In addition, researchers have access to better, more cost-effective computational power, more substantial and affordable storage via technologies such as the cloud, and faster networks. These changes in technology also allow broader participation in the scientific process, making it possible to successfully harness the public's participation through various citizen science activities (Newman et al., 2012).

Technology has changed the way scientists communicate and access information. Until recently, scientific knowledge dissemination was controlled by the major journal publishers within each field. This centralized means of communication was built upon a seventeenth-century publication model that was intended to make science more open at the time. However, a major disruption to this model occurred with the advent of the Internet, where scientists publish thoughts, ideas, results and conclusions openly and to everyone (Laakso et al., 2011). While trusted journals remain the primary medium for archiving and sharing peer-reviewed scientific knowledge, valuable information is also now available from these non-journal sources. These "gray literature" sources include reports, blogs, articles and various publications produced outside of the traditional commercial and academic publication workflows (Schöpfel & Prost, 2016). Non-traditional communication sources allow scientists to share initial results and lessons learned in real time and are challenging the prevailing belief that scientific results must be completely validated and verified before publication (De Roure et al., 2010). The details of results and analysis are also augmented by these new communication channels, which allow scientists to more thoroughly describe the development of algorithms and source code used to generate results.

Rapid technology advancements have increased the volume and variety of data by dramatically improving the instrumentation for observing and collecting data, the numerical models for simulations, and the processing abilities to efficiently analyze data. This rapid increase in the volume and velocity of data disrupts the scientist's traditional analytic workflows and the corresponding data management practices required for working with data. To complicate matters, a wide variety of heterogeneous data are now routinely used in scientific analyses. These data originate from a number of different sources including scientific instruments, models and unstructured data from non-traditional sources such as social media, online corpora of reports, and data gathered by citizen scientists Reichman et al., 2011). The exponential growth in data volume and complexity make science more reliant on complex computational platforms called cyberinfrastructures. Cyberinfrastructure, a term first used by the US National Science Foundation (NSF) (Cyberinfrastructure Council, 2007), is defined as an infrastructure consisting of computing systems, data storage systems, advanced instruments, data repositories, visualization environments and people all linked together by high-speed networks, making scholarly innovation and discoveries possible.

The increasing accessibility and availability of these data have opened opportunities to tackle complex, interdisciplinary science problems that span domain boundaries. Solving these complex problems requires collaboration across traditionally siloed scientific communities and the convergence of different types of expertise, knowledge and resources (Chesbrough, 2015). This focus on interdisciplinary research has led to a shift from individuals conducting research in isolation to a team approach with each member providing specialized expertise. In addition, in order to more effectively use sophisticated computation and data systems, science teams are frequently including programmers and computer scientists to help conduct analyses and to optimize algorithms for efficient analysis of large volumes of data.

The three open science drivers made the Human Genome Project (HGP), an exemplar of the open science movement, possible. The HGP was a 15 year program that set out to address a complex science problem: mapping and sequencing the human genome (Hood & Rowen, 2013; Watson, 1990). In order to address such a complex problem, the HGP was highly collaborative in nature and included international participation from 20 groups from the United States, the United Kingdom, Japan, France, Germany and China (International Human Genome Sequencing Consortium, 2001). Not only did the HGP promote international collaboration but, due to the complexity of the project, cross-disciplinary collaboration was also encouraged between computer scientists, mathematicians, engineers and biologists in order to make the needed advances in computational and mathematical approaches (Hood & Rowen, 2013; Collins et al, 2003). The generation of large data volumes, made possible due to a number of technological advances, required the open sharing of data. To support this open sharing ideal,the HGP drafted the 1996 Bermuda Principles, which committed the HGP laboratories to openly sharing DNA sequencing information on a daily basis (Cook-Deegan & McGuire, 2017). This open sharing of data led to new insights for biologists including the discovery of new information on 30 disease genes (Arias et al., 2015; Toronto International Data Release Workshop Authors, 2009). The HGP also upheld open-source principles by making the programs needed for analysis openly available (Hood & Rowen, 2013;). Most importantly, the HGP has had lasting scientific impacts on biology and medicine.The HGP led to the emergence of proteomics, a discipline focused on identifying and quantifying the proteins present in discrete biological compartments (Hood & Rowen, 2013;), has advanced scientists understanding of evolution, and initiated the comprehensive discovery and cataloguing of a 'parts list' of most human genes (Hood & Rowen, 2013). In addition, the HGP spawned a number of new scientific projects including the HapMap Project to catalogue human genetic variation (The International HapMap Consortium, 2005) and the ENCODE project (Encyclopedia Of DNA Elements) to understand the functional parts of the genome (Hood & Rowen, 2013; The ENCODE Project Consortium, 2011). The HGP's advances in open data sharing policies, open source code for analysis, technological advancements and an effective international collaboration model provide a benchmark for open science in the modern era.

3. Open Science Focus Areas

Figure 1 illustrates the three broad focus areas of open science: (1) increasing the accessibility to the scientific process and the corresponding body of knowledge; (2) making both the research process and knowledge sharing more efficient; and (3) understanding and assessing scientific impact through innovative new metrics.

3.1. Accessibility to Science

The first open science focus area concentrates on making science more accessible to both the scientific community and the general public. This focus area places an emphasis on providing access to science as a way of thinking but also on providing wider access to science as a body of knowledge.

3.1.1. Science as a Way of Thinking

The scientific method is a disciplined way of thinking and conducting research. In the past, access to the scientific research process was perceived to be limited to scientists with specialized degrees. While there have always been amateur scientists contributing to the scientific endeavor, the digital age has made it easier for more non-scientists to participate in research. Encouraging broader and more inclusive participation in all stages of the scientific research process is a key aspect of emerging open science activities. "Citizen science" is a commonly used term to describe the participation of non-scientists and amateurs in research (Fecher & Friesike, 2014). Citizen science activities leverage a volunteer workforce in a scientifically meaningful manner via activities such as systematically collecting data or analyzing data to discover interesting patterns.

Just as including amateurs in the scientific process is important, making research results understandable and comprehensible to the general public are equally important. Traditionally journal articles have been the primary method for communicating research results, but a journal's audience has typically been limited to the academic community. Other stakeholders, such as decision-makers, policymakers and the general public, want and need to understand scientific results; however by its nature science is complicated, and journal articles contain too much technical terminology for a layperson to understand.

To make science more accessible to all audiences, open science efforts offer new avenues, tools and formats of science communication beyond the journal model. These efforts include science writing that targets a broader community by deconstructing interesting yet complex research results into easy-to-understand pieces of information. Science blogs, such as Dan's Wild Wild Science Journal (Satterfield, n.d.), are one effective mechanism for communicating interesting scientific results. Social media platforms are also proving to be an effective communication tool for a broader audience. Astrophysicist Neil deGrasse Tyson has around 14 million followers on Twitter while physicist Brian Cox and biologist Richard Dawkins have around 3 million followers each. While blogs and social media are wide-reaching dissemination platforms, these tools are only as effective as the scientists who are willing to devote time and energy to creating these types of content. For this reason, there is a growing need for people, organizations or groups within the scientific community to establish themselves as creditable boundary spanners, or mediators, between scientists and nonscientists (Safford et al., 2017) for effective communication.

3.1.2. Science as a Body of Knowledge

Equitable access to the scientific body of knowledge is an essential dimension of open science. The scientific body of knowledge is the products of research and includes data, software, research publications and other supporting materials (Fecher & Friesike, 2014). Equitable access to these objects is enabled through open data policies, open source software principles and open access literature.

3.1.2.1. Open Data

Data drives the scientific process in two ways. First, data is a product of research activities and is a key component in the scientific body of knowledge. Second, data is analyzed for scientific insights and results. Since data is essential to the scientific process, open science efforts have focused on making data more openly available. Open data is data that may be accessed, used, and shared for any purpose without restrictions. Although the term 'open data' was initially used to describe open government data, the term has gained broader adoption as individuals and commercial institutions have recognized the value of providing open data. For government and commercial institutions, open data policies are essential for defining what data may be made openly available and for communicating to users how the data may be used. Open data policies define what data will be shared, with whom, at what price, and under what conditions the data can be reused or redistributed (Borowitz, 2017). Data sharing policies fall on a spectrum of openness with the most open data being made fully available either free of charge or at no more than the cost of reproduction (Group on Earth Observations, n.d.; Open Knowledge Foundation, 2020). On the other end of the spectrum, data may have limited or restricted access due to security concerns, the inclusion of personally identifiable information (PII), or licensing agreements often associated with commercial data purchases.

Open data benefits the open science movement in a number of ways. First, open data policies prevent duplicating the collection of data across organizations, freeing up resources to amass a more diverse array of data and making it possible to have a more comprehensive record of observations. For example, data exchange agreements between NASA and the European Space Agency (ESA) have made a virtual constellation of observations from the Landsat series and Sentinel-2 possible. Combining data from these two platforms increases the frequency of observations over land which is essential to land monitoring applications research. Second, open data policies significantly increase data use and reuse, especially when data is made freely available. The Landsat free and open data policy represents the epitome of a successful open data policy. After making Landsat data freely and openly available in 2008, the USGS saw a 20-fold increase in data downloads from 2009 to 2017 and a four-fold increase in the use of the data in the annual number of publications (Zhu et al., 2019). Providing open access to Landsat's long-term record of observations allowed scientists to move from only analyzing single images to conducting time-series analyses, enabling advances in a number of applications including monitoring land surface changes, tracking rates of change in shoreline erosion and measuring glacial fluctuations (Wulder et. al, 2012; Roy et. al, 2014; Kennedy et. al, 2014).

More broadly, there are a number of economic and societal benefits to open data. Remote sensing data have been of particular use to economists primarily because of its high spatial resolution, its wide geographic coverage and its ability to offer access to information unavailable by other means (Donaldson & Storeygard, 2016). Remote sensing data have been used in a variety of economic contexts including agriculture, infrastructure investments, tourism, resource availability and insurance (Donaldson & Storeygard, 2016). Additionally, there are a number of societal benefits to providing open data. Remote sensing data aids in disaster mitigation, response and recovery and is also a valuable input into monitoring conflicts, illegal activities, pollution events and the effects of policies on land use (Zhu et al., 2019; Donaldson & Storeygard, 2016).

3.1.2.2. Open Source Software

Software, in combination with data, acts as a tool for new knowledge discoveries and insights and often serves as a representation of knowledge in and of itself (Keyes & Taylor, 2011). However, software, unlike data, is protected by copyright, making the free use of software restricted unless the copyright owner has granted a license. Software is a broad term that applies to computer programs, applications, and source code that provide a certain level of utility to users or assist in producing a result (Committee on Best Practices for a Future Open Code Policy for NASA Space Science et al., 2018). In order to be considered open source, software must be made publicly available and include a software license that grants permissions for anyone to examine, use, change and distribute the source code for any purpose (Committee on Best Practices for a Future Open Code Policy for NASA Space Science et al., 2018). Source code without a license is considered "all rights reserved" and is therefore not available for free and open use.

Software is broadly categorized into two categories: scientific analysis tools and supporting infrastructure software. Scientific analysis tools include libraries, single-use code, analysis software, model and simulation software, modeling frameworks, and sensor and instrument data processing software (Committee on Best Practices for a Future Open Code Policy for NASA Space Science et al., 2018). Supporting infrastructure software includes any tools, services and infrastructure components developed that support science and provide services to manage data. Different software types will have unique software development approaches, but most software developed to support science should be assigned open source licensing.

Open source software culture is slowly gaining momentum within the sciences as the community recognizes the benefits and impacts of making software openly available. First, open source software encourages software reuse, the benefits of which include reducing the time working with data, lowering duplication of effort, enhancing the use of open data and ensuring the longevity of code (Committee on Best Practices for a Future Open Code Policy for NASA Space Science et al., 2018). Open source software also makes teamwork easier across organizations and allows scientists to have a larger collaborative network. Lastly, open code enables scientific and computational reproducibility and transparency. Scientific reproducibility is elusive without access to open code (Gil et al., 2016), and publications with only natural-language descriptions of methods, algorithms and code implementation are often insufficient for reproducibility (Committee on Best Practices for a Future Open Code Policy for NASA Space Science et al., 2018). Computational reproducibility is even more challenging to achieve, but new, forward-thinking approaches such as reusable research objects, executable publications, and containers are making computational reproducibility easier (Gil et al., 2016; Koop et al., 2011; Chard et al., 2020; Ton That et al., 2017).

3.1.2.3. Open Access Literature

Journal articles have been and continue to be a key mechanism for sharing scientific results with the broader community. However, access to journal articles has been limited by the prohibitive cost of accessing a large number of journals needed for research and by copyright restrictions that limit the freedom of sharing. Open access literature removes some of these limitations by making articles available digitally online, free of charge and free of most copyright and licensing restrictions (Suber, 2012). There are two types of open access literature: gold open access and green open access. Gold open access articles are delivered and distributed by journals. To be designated a gold article, authors simply submit a manuscript to an open-access journal (Suber, 2012). Green open access articles, however, are distributed by a repository. An author delivers a manuscript to an open access repository, an act also known as self-archiving, and the manuscript is required to be published in a journal that allows self-archiving (Suber, 2012). The flexibility of gold versus green open access makes it possible for authors to share manuscripts via a green repository, especially when a high prestige open access journal is not available for a given domain. There are a number of green repositories available including EarthArXiv (EarthArXiv, n.d.), OSFPreprints (Center for Open Science, n.d.), the Earth and Space Science Open Archive (ESSOAr, n.d.) and ArXiv (Cornell University, n.d.).

Journal publishers have adopted different publication approaches in order to make research articles openly accessible. For example, five of AGU's journals are inherently open access while sixteen other journals within AGU employ a hybrid model where a paper within a journal can be made open access if the article processing cost is borne by the author. AMS, SPIE, IEEE and Remote Sensing Society (GRSS) publications have also adopted different variations of this publication model to address the need for increasing access to research articles. In addition, new policies at some journals allow authors to self-archive in green repositories. These policies allow authors to offer manuscripts and other information openly to others so long as the repository is not-for-profit and encourages scientific engagement. Some publishers even provide a green repository for self-archiving, the most notable being AGU's ESSOAr.

3.2. Efficient Research Processes and Knowledge Dissemination

The second open science focus area considers how to make both the research process and scientific collaborations more efficient.

3.2.1. Cyberinfrastructure to Support Science

The shift in science toward data-intensive scientific discovery (Gray, 2009) has necessitated the need for new and better computation infrastructures to support science at scale. There have been a number of cyberinfrastructure projects (Droegemeier et al, 2005; Ludäscher et al, 2006; Nemani, 2012) that have focused on developing platforms with data management and analysis tools and services that make research efficient for different science domains. These projects free a researcher from the constraints of limited local resources by providing access to computing and storage resources along with a complementary suite of tools and services to enable large scale, data-intensive research, high throughput processing and modeling capabilities. These projects abstract away the data management and scaling complexity faced when working with large data volumes.

3.2.2. Collaborations

Large collaborative teams composed of individuals with different types of expertise are needed to solve increasingly complex and interdisciplinary scientific problems. These contributors play a number of roles in the scientific process including performing experiments, curating data, conducting analyses, developing software, validating results and conducting critical reviews (CASRAI, n.d.). Innovative new platforms are needed that seamlessly enable scientific collaboration from a number of geographically distributed contributors who are performing a variety of tasks. These effective collaborative platforms offer a number of key capabilities (Bartling & Friesike, 2014; Roure et al., 2008) including the ability to easily manage research objects, the incentivized sharing of those research objects, the capacity to be open and extensible to future technology changes, the ability to peer review scientific research objects (Himmelstein et. al, 2019) and the means to support actionable research beyond simply serving as an object repository. A number of collaborative open science platforms, such as myExperiment (De Roure et al., 2009), JetStream (Jetstream, n.d.) and GeneLab (NASA, n.d.-b; Berrios et al., 2020), support online collaborations in novel ways, along with science specific social networks, including ResearchGate and Mendeley (Nentwich & König, 2014), that allow researchers to connect and share journal articles.

3.3. Understanding Scientific Impact

The third open science focus area seeks to understand the broader implications of scientific research. As scientific research is shared more openly across a variety of platforms and to a number of different audiences, there is a need to comprehensively understand the impact of scientific contributions both to academia and to the broader public. Quantitative metrics, also known as impact measurements, offer one method for assessing scientific impact in the digital era. In the past twenty-five years, impact measurements have been limited to citation analysis of academic journal articles to assess scientific contributions (Fenner, 2014). While this constraint was initially necessary due to the physical printing of journal articles, the movement towards electronic publishing has eliminated this limitation. Citation analysis has offered some insights into scientific impact but the slow adoption of citations for some content like data and software (Fenner, 2014) means that many first-class research objects are not considered. Additionally, as scholarly workflows migrate to the web, other scientific impact measurements need to be considered such as article-level metrics, social network sharing metrics, and usage metrics (Fenner, 2014).

To address some of these impact measurement needs, the new discipline of altmetrics was established in 2010 to measure attention on the social web (Bar-Ilan et al., 2019). Altmetrics are collected for individual scientific outputs, such as a paper or dataset, and consider a variety of input variables including Mendeley document additions, social media shares or blog post references (Bar-Ilan et al., 2019). The input variables vary from organization to organization as does the weight each input variable receives when calculating an altmetric (Crotty, 2017) making altmetric values different from service to service. Altmetrics are calculated by a number of service providers including Altmetric.com (Altmetric, n.d.), PlumX (Plum Analytics, n.d.), ImpactStory (Our Research, n.d.) and Scienceopen.com (ScienceOpen, n.d.). However, some altmetric calculation methods, such as those from Altmetric.com, are proprietary, rendering these metrics fundamentally opposed to open science and open source principles.

4. Data Program's Role to Enable Open Science

The three open science focus areas can be encouraged and accelerated through thoughtful design (National Academies of Sciences, Engineering, and Medicine, 2018). Since modern science is primarily data-driven, data programs are in a unique position to design policies and systems that support and promote open science. This section details how data programs can enable all of the described open science focus areas and provides specific examples from NASA's Earth Science Data Systems (ESDS).

4.1. Accessibility to Science

4.1.1. Encouraging Science as a Way of Thinking

Data programs can support data collection activities that involve public participation. For example, ESDS's Citizen Science for Earth Systems Program (CSESP) focuses on developing and implementing projects that encourage the general public's contributions to advance Earth system science. NASA's citizen science projects have the same scientifically rigorous standards as any other science mission directorate project (SMD Science Management Council, 2018) yet offer citizens an opportunity to participate in the process. Data systems also support the legitimacy of citizen science efforts by ensuring these data are subject to key stewardship activities, including standardized documentation, metadata, file formats and quality assessments (Earth Science Data Systems, 2020a). The application of data stewardship processes ensures that these data are discoverable and usable to the broader community.

Data programs may also engage the public by systematically supporting scientific challenges, hackathons and other open events. These events benefit open science by harnessing the public's creativity and innovation and by enabling the broader public to learn about and engage with science data. Several examples of these types of events include NASA's International Space Apps Challenge (NASA, n.d.-c), the Copernicus hackathons (Copernicus, n.d.) and labeling events in platforms like Zooniverse (Zooniverse, n.d.) and the Sentinel-Hub Classification App (Sentinel Hub, n.d.). One archetypal example of a practical and collaborative approach to creating data challenges is the Climate Data Initiative (CDI) Innovation challenges. The Climate Data Initiative was an essential aspect of President Obama's Climate Action Plan (Office of the Press Secretary, 2014) that leveraged the federal government's extensive open data catalog to spur innovation and advance resilience to the impacts of climate change. The CDI Innovation challenges catalyzed new, data-driven solutions to help communities understand and build resilience to climate change. Ten innovation challenges were conducted as a part of the CDI and covered a wide range of topics from coastal flooding, disaster resiliency, food resiliency and the impacts of climate change on human health. Many public and private organizations participated in the innovation challenges, including NASA, NOAA, USGS, USDA, ESRI, Microsoft and Research Data Alliance. More recently, NASA hosted the Space Apps COVID-19 Challenge to solve problems surrounding the COVID-19 pandemic. Over 15,000 people from 150 countries used Earth observation data from NASA and its partner space agencies to show how satellite information can aid in the understanding of the COVID-19 outbreak on both global and local scales (Landau, 2020).

Lastly, data programs should make a concerted effort to increase public awareness of the value of the data collected to advance science and humanity. Similar to science blogs, systematic communication channels should be employed to craft data stories for the public. These stories create a connection between scientific data and how they impact people's lives. NASA's Earth Observatory website (NASA, n.d.-a) and the Land Processes Distributed Active Archive Center (LP DAAC) 's Data In Action (Land Processes Distributed Active Archive Center, 2020a) articles are two examples of content which communicate the broader importance of data to society. Data programs should also consider participating in working groups, such as the GEOValue community (GEOValue, n.d.), to better understand and promote Earth observation data's benefits to the broader community.

4.1.2. Enabling Access to Science as a Body of Knowledge

Data system programs support open access to knowledge by developing and implementing open data and open source software policies. These policies, such as the twenty-five-year-old ESDS open data policy, enable not only scientific research for a broad community of users but also make international collaboration possible. Open data policies empower partnerships, such as the Group on Earth Observations (GEO)’s partnership of 100 national governments, to occur. In turn, these open data policies enable the creation of new and innovative products that benefit the wider Earth observation community. For example, the Harmonized Landsat Sentinel (HLS) data product (Land Processes Distributed Active Archive Center, 2020b; Claverie et al., 2018) is being produced by NASA to support land applications and uses USGS's Landsat data and ESA's Sentinel 2 data. Data products like HLS are only possible when open data policies are in place.

Open data and software policies should be straightforward and easy to understand in order to benefit the user community and the scientists who are creating the data and software. Clear and unambiguous policies help users who need to understand any use or sharing constraints associated with data and software. Data programs can provide clarity to users by leveraging standard licenses for data and software instead of custom licenses, which may be interpreted in a number of ways. Straightforward policies also benefit scientists who are not subject matter experts in open data and open source software and licensing. A simple, permissive, and coherent open source software policy reduces the compliance burden on scientists and ensures a higher adoption rate of open source software principles.

Successful implementation of open policies by data programs may be achieved by providing clear and consistent guidelines for each step of the scientific data life cycle and by gaining the support of science program managers. Clear communication on policy expectations is essential and begins with solicitations and new research project requirements. Data programs can drive policies in the project formulation phase by including these policies in the data management plan (DMP) requirements for any project producing data of value to the broader community. Data programs should also consider supporting software management plans (SMPs) either as a stand-alone document or as a significant component within the DMP. Clear guidelines should also be provided regarding what data, software and code are expected to be openly shared. Educational resources, clear submission guidelines and well-formulated examples provided by the data program can ensure that DMPs, SMPs and the created research objects comply with open policies. While NASA's ESDS has well-documented guidelines for data management plans (Earth Science Data Systems, 2020d), software management plans are not currently required and are not a major component of DMPs. While ESDS does require that proposals include a plan for committing their software as OSS (Earth Science Data Systems, 2020c), an opportunity exists to more formally support SMPs within the program.

Once data and software are created, data programs can support open science by improving the discovery and use of crucial research objects through following strong data stewardship best practices and adopting the FAIR principles which state that data should be findable, accessible, interoperable, and reusable (Mons et al., 2017). These best practices include creating comprehensive metadata and assigning persistent identifiers and citations to all data and software. Comprehensive metadata should be provided for both data and software and should provide enough information to support understanding and reuse by a diverse community of users. While this approach to metadata is widely adopted for data, software should also be treated as a first-class research object. Requiring metadata, documentation, and preservation plans ensure that software will be treated as a first-class research object. Data programs should provide citations and DOIs for both data and software. While progress has been made in embracing data citation standards, there is a need for data programs to adopt software citation principles (Stodden et al., 2018). Meanwhile, authors of scientific papers are slow to adopt data and software citation best practices. Data programs can support citation by leveraging existing software citation standards and encouraging and supporting authors to use citations. While NASA's ESDS open source policy requires that all software source code developed through ESDS-funded research solicitations be designated, developed, and distributed to the public as OSS, the program still has a need to develop software metadata, citations and DOIs to enable broader discovery.

Finally, data programs should support access to research publications related to data and software. Digital library portals, such as the Astrophysics Data System (ADS) (Smithsonian Astrophysical Observatory, 2020), can serve not only as a primary entry point to relevant scientific publications but also as a linkage between journal articles and related data and code. Data programs should consider developing and maintaining infrastructures like the ADS which allows for the discovery of journal articles and also serves as a green repository for authors to upload preprints or journal articles that are free and open. The success of a data program-sponsored green repository can be magnified by incentivizing authors to participate in the open journal publication process. Authors should be encouraged to either publish in an open-access journal or to a journal that allows publishing to a green repository.

Accessibility to knowledge is only complete when data, software, documentation and publications are linked together and discoverable.

4.2. Enabling Efficient Research Processes and Knowledge Dissemination

4.2.1. From Data Systems to Enabling Collaborative Infrastructures

In most current data system architectures, the data archives are separate from computing resources. Any analysis requires data movement to either a user's machine or to some computing resource. Given the complexity of scientific data, multiple preprocessing steps are needed for the data to be analysis-ready. Furthermore, systematic data management is required to analyze and process data at scale. This overhead of data movement, management, and wrangling is accepted as part of the research overhead. Typically these data challenges are endured as an element of a researcher's science training.

de La Beaujardière, J. (2019) posed the question on how to enable "science at scale," such that researchers and other users can work with large, multisource data sets with minimal data management and wrangling overheads. Robinson et al. (2020) stated the need for infrastructures that facilitate the construction of a robust, scalable and adaptable data analysis pipeline. These infrastructures enable researchers to cope with the volume of data, provide effective user/data interfaces and visualizations, and utilize more powerful algorithms to extract more information from these datasets. These enabling infrastructures move researchers away from the drudgery of data management and wrangling, and back to intuitive and productive workflows thus accelerating scientific progress. These enabling infrastructures need to support data processing, data storage for archive and rapid access, data analytical methods for information extraction and knowledge generation, multiple modalities for user interaction, and collaboration. Furthermore, these infrastructures need to support various end-users including scientists, students, engineers and decision-makers (Yang, C. et al., 2019). While there have been numerous efforts to build cyberinfrastructures, efforts are needed that go beyond such approaches in that they couple infrastructures with research software and research practices to enable more efficient Earth system science research (Bandaragoda et al., 2019).

Cloud computing is quickly becoming a viable approach for building such infrastructures that move data out of institutional silos and into a common computational platform (de La Beaujardière, J. 2019). As a new computing paradigm, cloud computing delivers scalable, on-demand, pay-as-you-go access to a pool of computing resources. These cloud technologies provide the building blocks needed to make it easier, more efficient and more economical to configure data analysis platforms for scaled computing and reduced time to analysis. Therefore, data systems need to be reimagined to be cloud native and more fully integrated with analysis platforms.

A conceptual cloud-native data system that supports new enabling cyberinfrastructures is depicted in the conceptual figure 2 below (Bugbee et al., 2020). The data lake serves as a central repository of different data. The core set of data services cover scientific data stewardship functions such as metadata generation, data ingest, reformatting, documentation and data publication, and services required by a user to discover, visualize and access the data. The data services also allow direct access to the data files. Structured data stores within this data ecosystem can be a stand-alone component or instantiated when needed, as envisioned in an analytics optimized data store (Ramachandran et al., 2019). Implementation of structured data stores can range from the use of new database technologies such as Rasdaman (Baumann et al., 2013) to data cubes (Giuliani et al., 2019) to software frameworks like Pangeo (Guillaume et al., 2019), which utilize cloud-optimized formats like Zarr and Dask for parallelizing processing tasks under the hood. These structured data stores minimize the data wrangling burden on the end-users and allow fast reads on the data.

The cloud's virtualization capability enables users to package their code as containers and run it on the data residing in the data lake. The ability to perform analysis either using a structured data store or by running containers on the data lake minimizes the need to download data. The data spaces component provides a private personalized workspace where users can curate and integrate different data. Data spaces coupled to containers and workflows facilitate sending code to the data and running analysis at scale. Collaboration within data spaces, when provided, allows the sharing of both code and data. As data is transformed within the ecosystem's different components, core data services need to include additional data stewardship activities such as documenting data and lineage. A standard set of APIs serves as a connector between various data system components, and new client tools and services can utilize these APIs.

Data programs should invest in developing open-source software that supports implementing enabling collaborative cyberinfrastructures by leveraging current advances in technology such as cloud computing. The Joint ESA-NASA Multi-Mission Algorithm and Analysis Platform (MAAP) is one such example. The MAAP is a collaborative, cloud-based, open science platform dedicated to the biomass community's unique research needs (Bugbee et al., 2020). The MAAP, as a virtual platform, provides a new approach to accessing, sharing, analyzing and processing data. The MAAP's users have seamless access to open airborne, spaceborne and field ESA and NASA data for biomass mapping. The MAAP provides the ability to upscale a user's algorithms, typically implemented as a Jupyter notebook, from small regions of interest to a global scale with minimal effort. The MAAP platform also enables users to collaborate on end-to-end calibration and validation of algorithms for generating higher-level science products. Pangeo is another example of a community platform, funded by NSF and NASA, that supports promoting open, reproducible and scalable science (Guillaume et al, 2019).

Data programs should also invest in collaborative, open source tools that support open science principles within the data program itself. Tools such as ESDS's Algorithm Publication Tool (APT) make it easier for scientists to collaboratively write Algorithm Theoretical Basis Documents (ATBDs)(Bugbee et al., 2020). ATBDs describe the physical theory, mathematical procedures and assumptions made for the algorithms that create higher level data products. These documents reinforce the data program's commitments to open science by supporting reproducibility and transparency of data products. Providing a tool like the APT makes it easier for scientists to collaborate on these documents and, in the end, makes the data program more open. The availability, adoption and use of such software to provide new enabling cyberinfrastructures are needed to make research processes efficient and accelerate knowledge dissemination.

4.3. Understanding Scientific Impact Through Data System Measurements

Data programs should adopt or develop impact measurements for all first-class research objects within the data system, including data, software, documentation, services and users. Impact measurements help data programs understand the value and use of data, software, services and information not just within the scientific community but more broadly. A greater understanding of these objects' impact helps inform future decisions about new projects, new technological infrastructure needs, and new data stewardship requirements. Insights from data system impact measurements also help scientists understand and receive credit for reused data and software. A secondary benefit to these measurements is a better understanding of the effectiveness of internal data system processes including data stewardship best practices. Lastly, data system impact measurements may enhance the open science process by serving as a metric or weight for discovering new research objects. Combining impact metrics with traditional search techniques enables discovery by highlighting research objects of interest.

While most data systems, including NASA's ESDS, collect data metrics (Earth Science Data Systems, 2020b), quantifying data and software impact metrics in the research community has been difficult in the past due to inconsistent or non-existent citation practices. Emerging methods, such as altmetrics and machine learning techniques, make it possible to understand the impacts of data, software and other information. For example, leveraging machine learning techniques across open access repositories and other unstructured data assets provides an alternative way to assess data and software impacts. Additionally, a new type of altmetrics for data systems that combine data, software, code and documentation usage may provide another means of assessing impact.

While traditional altmetrics have focused on journal articles as the primary research object of interest, the same approach may not be effective for data and software. Data system altmetrics should include already established input variables such as social media shares but should also explore other variables of interest, including measures of data stewardship such as metadata quality scores or data stewardship maturity matrix assessment scores (Peng et al., 2019). In addition, data system impact metrics should consider both data usage and data citations since there is some data usage by groups that never publish results in a peer-reviewed paper (Lowenberg et al., 2019). Similarly, software usage should be assessed by how widely a piece of code is supported by the community through measures such as forks on Github and incorporation into other tools and workflows. Lastly, documentation, such as algorithm theoretical basis documents (ATBDs), data guides, data recipes and blogs should be treated as first-class research objects within the data system. Assigning DOIs and citations to these key documents will make it possible to track these objects' impact and inclusion as a variable for data system altmetrics. Establishing linkages between data, software and documentation, similar to those seen in Scholix Link Information Packages (Lowenberg et al., 2019), are a comprehensive way of measuring impact across the data system.

Most importantly, the data system altmetrics should conform to open science and open-source principles. The methodology used to generate altmetrics should be transparent and reproducible, while the code used to create altmetrics should be open-sourced. Ensuring data system altmetrics are open also makes community buy-in and agreement possible.

5. Challenges for Data Programs

There are a number of challenges facing data programs that must be considered when supporting the open science paradigm. These challenges require creative solutions along with community engagement and discussion and are described below by specific focus areas.

5.1. Accessibility to Science

Greater engagement with the public is an important aspect of open science yet there are still hurdles to be overcome to support this engagement. For example, while there is a growing interest in supporting citizen scientist activities, there is some reluctance to accept citizen science data as scientifically legitimate from both within the science community and even by some data centers. Overcoming these biases will be critical to ensure the long term adoption and viability of citizen science data. In addition, many scientists and data centers are adverse to committing limited resources in order to effectively communicate data value to the wider public. To encourage more scientists to become these types of boundary spanners, the scientific community will need to evolve to value these contributions to science. Data programs and scientific programs can foster acceptance by providing funding for these types of roles and creating solicitations that emphasize science communication to the broader public.

Providing open access to data makes it possible for data to be reused for scientific benefit but also poses some issues for data programs. First, making data openly available introduces conditions for data misuse. Individual users may unwittingly use data in inappropriate ways or, more insidiously, use data to misrepresent facts. For example, Earth observation data was recently misused or misrepresented in maps and images generated during the Australian fire crisis (BBC, 2020). Similarly, some third-party organizations bulk download open data in order to redistribute the data. The level of data stewardship and understanding provided by these third party distributors is typically lower than those of the original data providers. This disparity can lead to data misunderstanding and misuse in platforms such as Google Earth Engine, where users do not always understand the need to use quality flags or the caveats surrounding data use. Lastly, requiring researchers to make data openly available is sometimes seen as burdensome and a distraction from research itself. Many scientists acknowledge that they are holding "dark data, or data that has never been published or otherwise made available to the rest of the scientific community" (Heidorn, 2008). With little professional rewards for sharing these data and a lack of time and resources, scientists are often unwilling to put any effort into making these data openly available. In order to increase open data sharing, data programs need to develop solutions that minimize the burden on researchers to share data and also foster incentives and credits for sharing.

Open source software presents similar problems for both researchers and data programs. Similar to open data requirements, a move towards open source software introduces extra requirements for scientists that is often viewed as distracting (Committee on Best Practices for a Future Open Code Policy for NASA Space Science et al., 2018). Scientists are often not experts in the open source process and are confused by the lack of consensus as to where to share code and which license to select for code. Assigning an appropriate license to software is confusing and intimidating to a novice code developer and is often overwhelming for scientists who want to spend time pursuing science and not learning copyright laws. This confusion is compounded by the fact that the use of standard open-source licenses may be limited by the constraints imposed by government mandates. In addition to the licensing, there is currently minimal guidance on which code repositories should be used for sharing software and code. Some source code is individually shared using code repositories such as Github while some organizations leverage an institutional Github/Gitlab repository instance. Official repositories, such as the IEEE Remote Sensing Code Library (IEEE, 2020), may alleviate this issue but options are currently limited. Data programs may need to consider providing clear and easy-to-understand guidelines on the open source software process to promote success.

While open access publication primarily poses challenges for researchers, data programs have a need to provide free and open access to journal articles related to data and software. Data programs may address this need by providing a green repository for open access to relevant journals and by working with scientists to ensure data articles are published in journals which support self-publication.

5.2. Enabling Efficient Research Processes and Knowledge Dissemination

Developing enabling infrastructures that support both collaboration and analysis at scale requires a fundamental transformation in existing data systems. Data programs must invest in and evolve data systems towards infrastructures that facilitate analysis. This data system evolution will require data centers to move beyond simply functioning as a data archive to instead serving as knowledge centers for the scientific community. This evolution will not be easy due to the sunk costs of existing processing, ingest, archive and distribution systems along with the need to retrain existing staff for new roles and responsibilities. Yet, data programs may view adopting cloud technologies as an opportunity to design and build new data analysis platforms that function as a cohesive and interdisciplinary ecosystem rather than individual, stand-alone data silos.

For data programs, there is a risk of building new infrastructure solutions that do not meet the scientific community's needs or have low adoption rates that do not justify the cost of implementation. The 'Build it, and they will come' type of tool development mentality always runs a risk of low adoption if the potential users are not engaged at the beginning of a project (Cutcher-Gershenfeld et al., 2016). This risk is especially true for collaboration tools which need to be designed to help scientists do what they are already doing, not what the tool developer feels scientists should be doing. Data programs may mitigate this risk by including scientific stakeholders early and throughout the infrastructure development process.

Similarly, while cloud computing has the potential to be a transformational technology for improving research processes and collaboration, cloud computing is not a panacea. For data programs, developing cloud-native tools and learning to utilize the cloud platform effectively in a cost-optimal manner is a non-trivial task. Advances in cloud platforms happen very rapidly, with new services added every week. These rapid advances place an extra burden on code maintenance and refactoring. Clearly, the cost of managing and sustaining these new types of cloud-based infrastructures will require the development of new sustainable business models and best practices. Cloud-based collaborative infrastructures may also require advances in data stewardship practices. As cloud computing enables different groups to build data sharing and analysis platforms quickly, there is a risk that these collaboration platforms will lead to data lakes turning into swamps filled with data of dubious quality (Sicular, 2016). These platforms will require systematic, semi-automated data governance and management plans in order to prevent the degradation of the data lake into a data swamp.

Cloud computing poses some challenges for researchers as well. Cloud computing still does not solve equity issues due to the disparity in bandwidth and the lack of needed technological infrastructures. The lack of access to the internet, high bandwidths, or necessary computing resources means that not everyone has equal access to open data and information. Similarly, while cloud computing reduces the initial cost needed to run analyses at scale, funding is still required to run cloud computing. The need for funding is particularly pressing as science adopts AI/ML techniques to build models which require computing at a scale that only a few organizations can afford. Finally, cloud capacity building is needed to assist researchers and end-users to effectively utilize cloud platform capabilities.

5.3. Impact Measurements

There are still several obstacles to be addressed in order for data programs to calculate data system altmetrics effectively. The biggest hurdle centers around the need to accurately understand and calculate data usage beyond the immediate data system itself. Data programs collect download metrics, and data and software citations in journal articles may serve as one measure of usage. However, data and code may be used for applications and in decision-making processes that are never published in a peer-reviewed journal. Also, as more third-party vendors harvest open data and make it available for use on alternative platforms, data programs will need to decide how to account for the varied usage of valuable data on these platforms.

More broadly, the acceptance and adoption of data system altmetrics is not assured. While data programs can take steps to ensure wider adoption, such as leveraging these metrics when reviewing new proposals, there is no guarantee that almetrics will be deemed of value by the community. Even if altmetrics are adopted, there are still limitations associated with collecting these metrics. Data, software and research may receive attention for the wrong reasons, thus misrepresenting the metrics and, in the end, the value of an object. In addition, there is a risk that these metrics are subject to self-promotion, gaming, and the constantly changing nature of the web (Fenner, 2014). Variables such as the career stage of an object creator, the specific discipline in which an object is created, the number and mix of co-authors and whether an object is cross-disciplinary in nature makes metrics comparisons problematic at best (Kurtz and Henneken, 2016). Effectively calculating impact metrics will require careful thought by data programs. However, if fairly and mindfully adopted, these metrics have the potential to provide greater insights into the impacts of data more broadly.

6. Conclusions

The open science movement continues to gain momentum as both scientists and institutions adopt many of the principles and ideas described in this paper. Open science beckons with the promise of more collaborative and efficient research, a more educated and engaged public, and scientific results that are reproducible and easier to understand. No doubt, there will be issues as the open science paradigm shift continues to expand. Data and compute equity will always be a challenge until reliable internet access is available to the majority. In addition, there are risks associated with both guaranteeing data authenticity and increasing data misuse. However, these issues should not keep us, as a community, from moving open science forward.

In order to move the open science paradigm forward, contributions are needed from both supporting institutions such as data programs and journal publishers but also individuals from the research community. As highlighted in this paper, data programs need to acknowledge the major role they play in supporting open science through the development of open policies, investment in innovative and collaborative infrastructures, and the promotion of cultural change.

Data programs should support data collection activities that involve public participation and help legitimize such citizen science efforts by enforcing standard scientific stewardship activities. The application of scientific data stewardship processes will ensure both usability and trust of the collected data. Engaging the public via scientific challenges and hackathons will enable data programs to harness the public's creative and innovative engagement with science data and increase public awareness of the scientific data's value. Data system programs should support open access to knowledge by developing and implementing open data and open source software policies and improving the discovery and use of crucial research objects. Knowledge is open only when data, software, documentation and publications are linked and discoverable. Data programs should invest in developing open-source software necessary for the next generation of enabling cyberinfrastructures that remove the drudgery of data management and wrangling and allow the use of open science tools to transform science. Finally, data programs should adopt or develop impact measurements for all first-class research objects within the data system, including data, software, documentation, services and users. Such measures will help guide future investments both in science as well as the data and information systems.

Individual researchers, on the other hand, can be advocates for open science by adopting a number of best practices. First, researchers should make their data available in an open repository and in a non-proprietary, standardized format. Whenever possible, they should create a digital object identifier (DOI) for their data and be sure to provide clear licensing information along with use constraints about the data. Second, software and code should be made open source through a community adopted license that is as permissive as possible to encourage reuse. In addition, any libraries used should also open sourced and follow software development best practices, such as rigorous version controls. Third, researchers should contribute back by supporting the community development of open source software, libraries, code and tools that are used by the wider community. When appropriate, code should be open to community development and feedback as well. Fourth, peer-reviewed articles should be published to gold journals when possible. If a reputable gold journal is not available for a particular domain, researchers should make sure their publishers allow self-publishing in a green repository. Fifth, the public should be actively engaged by starting or supporting science blogs, citizen science projects, hack-a-thons or sharing results on social media. Finally, researchers should cite data, software and documentation whenever possible but especially in journal articles.

Together individuals and institutions will usher in a new and more open age of scientific research that will not only benefit science itself but will bring further advances to society as a whole.

Acknowledgements

Authors would like to thank Derek Koehl for his assistance in editing and improving the manuscript. We also thank all our colleagues within the Earth Science Data Systems Program NASA who provided insight and expertise that greatly assisted this research.

References

Altmetric. (n.d.). *Discover the attention surrounding your research*. Altmetric. Retrieved October 22, 2020, from https://www.altmetric.com/

Arias, J. J., Pham-Kanter, G., & Campbell, E. G. (2015). The growth and gaps of genetic data sharing policies in the United States. *Journal of Law and the Biosciences*, *2*(1), 56–68. https://doi.org/10.1093/jlb/lsu032

Bandaragoda, C., Castronova, A., Istanbulluoglu, E., Strauch, R., Nudurupati, S. S., Phuong, J., Adams, J. M., Gasparini, N. M., Barnhart, K., Hutton, E. W. H., Hobley, D. E. J., Lyons, N. J., Tucker, G. E., Tarboton, D. G., Idaszak, R., & Wang, S. (2019). Enabling Collaborative Numerical Modeling in Earth Sciences using Knowledge Infrastructure. *Environmental Modelling and Software*, *120*, 104424. https://doi.org/10.1016/j.envsoft.2019.03.020

Bar-Ilan, J., Halevi, G., & Milojević, S. (2019). Differences between Altmetric Data Sources – A Case Study. *Journal of Altmetrics*, *2*(1), 1. https://doi.org/10.29024/joa.4

Bartling, S., & Friesike, S. (2014). Towards Another Scientific Revolution. In S. Bartling & S. Friesike (Eds.), *Opening Science* (pp. 3–15). Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_1

Baumann, P., Dumitru, A. M., & Merticariu, V. (2013). The Array Database That Is Not a Database: File Based Array Query Answering in Rasdaman. In M. A. Nascimento, T. Sellis, R. Cheng, J. Sander, Y. Zheng, H.-P. Kriegel, M. Renz, & C. Sengstock (Eds.), *Advances in Spatial and Temporal Databases* (pp. 478–483). Springer Berlin Heidelberg.

BBC. (2020). Australia fires: Misleading maps and pictures go viral. *BBC News*. https://www.bbc.com/news/blogs-trending-51020564

Berrios, D. C., Galazka, J., Grigorev, K., Gebre, S., & Costes, S. V. (2020). NASA GeneLab: Interfaces for the exploration of space omics data. *Nucleic Acids Research*, gkaa887. https://doi.org/10.1093/nar/gkaa887

Borowitz, M. (2017). *Open space: The global effort for open access to environmental satellite data*. The MIT Press.

Bugbee, K., Lynnes, C., Ramachandran, R., Maskey, M., Barciauskas, A., Kaulfus, A., Ton-That, D.-H., Virts, K., & Markert, K. (2020). Advancing open science through innovativbe data system solution: The joint ESA-NASA Multi-mission Algorithm and Analysis Platform (MAAP)'S data ecosystem. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*.

CASRAI. (n.d.). *CRediT – Contributor Roles Taxonomy*. CASRAI. Retrieved November 6, 2020, from https://casrai.org/credit/

Center for Open Science. (n.d.). *OSF*. OSFHome. Retrieved October 9, 2020, from https://osf.io/

Chard, K., Gaffney, N., Hategan, M., Kowalik, K., Ludäscher, B., McPhillips, T., Nabrzyski, J., Stodden, V., Taylor, I., Thelen, T., Turk, M. J., & Willis, C. (2020). Toward enabling reproducibility for data-intensive research sing the whole tale platform. *Advances in Parallel Computing*, *36*, 766–778. https://doi.org/10.3233/APC200107

Chesbrough, H. (2015). *From open science to open innovation*. Institute for Innovation and Knowledge Management, ESADE.

Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J.-C., Skakun, S. V., & Justice, C. (2018). The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment*, *219*, 145–161. https://doi.org/10.1016/j.rse.2018.09.002

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: Lessons from Large-Scale Biology. *Science*, *300*(5617), 286–290. https://doi.org/10.1126/science.1084564

Committee on Best Practices for a Future Open Code Policy for NASA Space Science, Space Studies Board, & Division on Engineering and Physical Sciences. (2018). *Open Source Software Policy Options for NASA Earth and Space Sciences* (p. 25217). National Academies of Sciences, Engineering, and Medicine. https://doi.org/10.17226/25217

Cook-Deegan, R., & McGuire, A. L. (2017). Moving beyond Bermuda: Sharing data to build a medical information commons. *Genome Research*, *27*(6), 897–901. https://doi.org/10.1101/gr.216911.116

Copernicus. (n.d.). *Copernicus Hackathos*. EU. Retrieved October 9, 2020, from https://hackathons.copernicus.eu/

Cornell University. (n.d.). *ArXiv.org e-Print archive*. Retrieved October 27, 2020, from www.arxiv.org

Crotty, D. (2017). Altmetrics. *European Heart Journal*, *38*(35), 2647–2648. https://doi.org/10.1093/eurheartj/ehx447

Cutcher-Gershenfeld, J., Baker, K. S., Berente, N., Carter, D. R., DeChurch, L. A., Flint, C. C., Gershenfeld, G., Haberman, M., King, J. L., Kirkpatrick, C., Knight, E., Lawrence, B., Lewis, S., Lenhardt, W. C., Lopez, P., Mayernik, M. S., McElroy, C., Mittleman, B., Nichol, V., … Zaslavsky, I. (2016). Build it, but will they come? A geoscience cyberinfrastructure baseline analysis. *Data Science Journal*, *15*. https://doi.org/10.5334/dsj-2016-008

Cyberinfrastructure Council. (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. National Science Foundation.

de La Beaujardière, J. (2019). A Geodata Fabric for the 21st Century. *EOS*, *100*. https://doi.org/10.1029/2019EO136386

De Roure, D., Goble, C., Aleksejevs, S., Bechhofer, S., Bhagat, J., Cruickshank, D., Fisher, P., Hull, D., Michaelides, D., Newman, D., Procter, R., Lin, Y., & Poschen, M. (2010). Towards open science: The myExperiment approach. *Concurrency and Computation: Practice and Experience*, *22*(17), 2335–2353. https://doi.org/10.1002/cpe.1601

De Roure, D., Goble, C., & Stevens, R. (2009). The design and realisation of the Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems*, *25*(5), 561–567. https://doi.org/10.1016/j.future.2008.06.010

Donaldson, D., & Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, *30*(4), 171–198. https://doi.org/10.1257/jep.30.4.171

Droegemeier, K. K., Gannon, D., Reed, D., Plale, B., Alameda, J., Baltzer, T., Brewster, K., Clark, R., Domenico, B., Graves, S., Joseph, E., Morris, V., Murray, D., Ramachandran, R., Ramamurthy, M., Ramakrishnan, L., Rushing, J., Weber, D., Wilhelmson, R., … Yalda, S. (2005). Service-Oriented Environments in Research and Education for Dynamically Interacting with Mesoscale Weather. *IEEE Computing in Science & Engineering*, *7*(6), 24–32.

Earth Science Data Systems. (2020a). *NASA ESDS Citizen Science Data Working Group White Paper* (1.0). NASA. https://cdn.earthdata.nasa.gov/conduit/upload/14273/CSDWG-White-Paper.pdf

Earth Science Data Systems. (2020b, July 14). *EOSDIS Annual Metrics Reports*. NASA Earth Science Data Systems Program. https://earthdata.nasa.gov/eosdis/system-performance/eosdis-annual-metrics-reports

Earth Science Data Systems. (2020c, September 24). *ESDS Open Source Software Policy*. NASA Earth Science Data Systems Program. https://earthdata.nasa.gov/collaborate/open-data-services-and-software/esds-open-source-policy

Earth Science Data Systems. (2020d, October 20). *Data Management Plan Guidance*. NASA Earth Science Data Systems Program. https://earthdata.nasa.gov/esds

EarthArXiv. (n.d.). *EarthArXiv*. EarthArXiv. Retrieved October 9, 2020, from https://eartharxiv.org/

ESSOAr. (n.d.). *Earth and Space Science Open Archive*. ESSOAr. Retrieved October 9, 2020, from https://www.essoar.org/

Fecher, B., & Friesike, S. (2014). Open Science: One Term, Five School of Thoughts. In *Opening Science*. Springer.

Fenner, M. (2014). Altmetrics and Other Novel Measures for Scientific Impact. In Sonke Bartling & S. Friesike (Eds.), *Opening science* (pp. 179–189). Springer. https://doi.org/10.1007/978-3-319-00026-8

Friesike, S., Widenmayer, B., Gassmann, O., & Schildhauer, T. (2015). Opening science: Towards an agenda of open science in academia and industry. *The Journal of Technology Transfer*, *40*(4), 581–601. https://doi.org/10.1007/s10961-014-9375-6

GEOValue. (n.d.). *Geo Value*. GEO Value. www.geovalue.org

Gil, Y., David, C. H., Demir, I., Essawy, B. T., Fulweiler, R. W., Goodall, J. L., Karlstrom, L., Lee, H., Mills, H. J., Oh, J., Pierce, S. A., Pope, A., Tzeng, M. W., Villamizar, S. R., & Yu, X. (2016). Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science*, *3*(10), 388–415. https://doi.org/10.1002/2015EA000136

Giuliani, G.; Camara, G.; Killough, B.; Minchin, S. (2019). Earth Observation Open Science: Enhancing Reproducible Science Using Data Cubes. *Data*, *4*, 147.

Gray, J. (2009). A Transformed Scientific Method. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.

Group on Earth Observations. (n.d.). *GEO Data Sharing Principles Implementation*. Group on Earth Observations. Retrieved October 9, 2020, from https://www.earthobservations.org/geoss_dsp.shtml

Guillaume, E.-B., Abernathey, R., Hamman, J., Ponte, A., & Rath, W. (2019). The Pangeo Big Data Ecosystem and its use at CNES. In *Big Data from Space (BiDS'19).... Turning Data into insights...*https://archimer.ifremer.fr/doc/00503/61441/

Hansen, K. (2012, July). Five outstanding questions in Earth science. *Earth*, *57*(7), 40–45.

Himmelstein, D. S., Rubinetti, V., Slochower, D. R., Hu, D., Malladi, V. S., Greene, C. S., & Gitter, A. (2019). Open collaborative writing with Manubot. *PLOS Computational Biology*, *15*(6), e1007128. https://doi.org/10.1371/journal.pcbi.1007128

Hood, L., & Rowen, L. (2013). The human genome project: Big science transforms biology and medicine. *Genome Medicine*, *5*(9), 79. https://doi.org/10.1186/gm483

IEEE. (2020). *Remote Sensing Code Library*. IEEE. http://www.grss-ieee.org/publication-category/rscl/

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Jetstream. (n.d.). *Jetstream: A National Science and Engineering Cloud*. Jetstream. Retrieved November 10, 2020, from https://jetstream-cloud.org

Kennedy, R. E., Andréfouët, S., Cohen, W. B., Gómez, C., Griffiths, P., Hais, M., Healey, S. P., Helmer, E. H., Hostert, P., Lyons, M. B., Meigs, G. W., Pflugmacher, D., Phinn, S. R., Powell, S. L., Scarth, P., Sen, S., Schroeder, T. A., Schneider, A., Sonnenschein, R., … Zhu, Z. (2014). Bringing an ecological view of change to Landsat-based remote sensing. *Frontiers in Ecology and the Environment*, *12*(6), 339–346. https://doi.org/10.1890/130066

Keyes, D., & Taylor, V. (2011). *National Science Foundation Advisory Committee for CyberInfrastructure Task Force on Software for Science and Engineering* [Final

Report]. National Science Foundation. https://www.nsf.gov/cise/oac/taskforces/TaskForceReport_Software.pdf

Koop, D., Santos, E., Mates, P., Vo, H. T., Bonnet, P., Bauer, B., Surer, B., Troyer, M., Williams, D. N., Tohline, J. E., Freire, J., & Silva, C. T. (2011). A provenance-based infrastructure to support the life cycle of executable papers. *Procedia Computer Science*, *4*, 648–657. https://doi.org/10.1016/j.procs.2011.04.068

Kurtz, M. J., & Henneken, E. A. (n.d.). *Measuring Metrics--A forty year longitudinal cross--validation of citations, downloads, and peer review in Astrophysics*.

Laakso, M., Welling, P., Bukvova, H., Nyman, L., Björk, B.-C., & Hedlund, T. (2011). The Development of Open Access Journal Publishing from 1993 to 2009. *PLoS ONE*, *6*(6), e20961. https://doi.org/10.1371/journal.pone.0020961

Land Processes Distributed Active Archive Center. (2020a). *Data in Action*. LPDAAC. https://lpdaac.usgs.gov/resources/data-action/

Land Processes Distributed Active Archive Center. (2020b). *Harmonized Landsat-Sentinel 2 (HLS) Overview*. LPDAAC. https://lpdaac.usgs.gov/data/get-started-data/collection-overview/missions/harmonized-landsat-sentinel-2-hls-overview/

Landau, E. (2020, September 18). *NASA Space Apps COVID-19 Challenge Winners Share Stories of Innovation*. NASA. https://www.nasa.gov/feature/nasa-space-apps-covid-19-challenge-winners-share-stories-of-innovation

Lowenberg, D., Chodacki, J., Fenner, M., Kemp, J., & Jones, M. B. (2019). *Open Data Metrics: Lighting the Fire*(Version 1) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.3525349

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E. A., Tao, J., & Zhao, Y. (2006). Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, *18*(10), 1039–1065. https://doi.org/10.1002/cpe.994

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H., & Yarkoni, T. (2016). How open science helps researchers succeed. *ELife*, *5*, e16800. https://doi.org/10.7554/eLife.16800

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, *37*(1), 49–56. https://doi.org/10.3233/ISU-170824

Murphy, M. C., Mejia, A. F., Mejia, J., Yan, X., Cheryan, S., Dasgupta, N., Destin, M., Fryberg, S. A., Garcia, J. A., Haines, E. L., Harackiewicz, J. M., Ledgerwood, A., Moss-Racusin, C. A., Park, L. E., Perry, S. P., Ratliff, K. A., Rattan, A., Sanchez, D. T., Savani, K., … Pestilli, F. (2020). Open science, communal culture, and women's participation in the movement to improve science. *Proceedings of the National Academy of Sciences*, *117*(39), 24154. https://doi.org/10.1073/pnas.1921320117

NASA. (n.d.-a). *Earth Observatory*. NASA. Retrieved October 9, 2020, from https://earthobservatory.nasa.gov/

NASA. (n.d.-b). *NASA GeneLab: Open Science for Life in Space*. GeneLab. Retrieved November 6, 2020, from https://genelab.nasa.gov

NASA. (n.d.-c). *NASA International Space Apps Challenge*. Space Apps Challenge. https://www.spaceappschallenge.org/

National Academies of Sciences, Engineering, and Medicine (Ed.). (2018). *Open science by design: Realizing a vision for 21st century research*. The National Academies Press.

National Academies of Sciences, Engineering, & Medicine 2018. (2018). Open Source Software Policy Options for NASA Earth and Space Sciences. In *National Academies of Sciences, Engineering, and Medicine 2018*. The National Academies Press. https://doi.org/10.17226/25217

Nemani, R. (2012). NASA EARTH EXCHANGE: NEXT GENERATION EARTH SCIENCE COLLABORATIVE. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, *XXXVIII-8/*, 17. https://doi.org/10.5194/isprsarchives-XXXVIII-8-W20-17-2011

Nentwich, M., & König, R. (2014). Academia Goes Facebook? The Potential of Social Network Sites in the Scholarly Realm. In Sönke Bartling & S. Friesike (Eds.), *Opening Science* (pp. 107–124). Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8_7

Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., & Crowston, K. (2012). The future of citizen science: Emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, *10*(6), 298–304. https://doi.org/10.1890/110294

Office of the Press Secretary. (2014, March 19). *FACT SHEET: The President's Climate Data Initiative: Empowering America's Communities to Prepare for the Effects of Climate Change*. The White House. https://obamawhitehouse.archives.gov/the-press-office/2014/03/19/fact-sheet-president-s-climate-data-initiative-empowering-america-s-comm

Open Knowledge Foundation. (2020). *What is open?* Open Knowledge Foundation. https://okfn.org/opendata/

Our Research. (n.d.). *ImpactStory*. ImpactStory. Retrieved October 9, 2020, from www.impactstory.org

P. Bryan Heidorn. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, *57*(2), 280–299. https://doi.org/10.1353/lib.0.0036

Peng, G., Milan, A., Ritchey, N. A., Partee, R. P., Zinn, S., McQuinn, E., Casey, K. S., Lemieux, P., Ionin, R., Jones, P., Jakositz, A., & Collins, D. (2019). Practical application of a data stewardship maturity matrix for the NOAA onestop project. *Data Science Journal*, *18*(1). https://doi.org/10.5334/dsj-2019-041

Plum Analytics. (n.d.). *About PlumX Metrics*. Plum Analytics. Retrieved October 9, 2020, from https://plumanalytics.com/learn/about-metrics/

Rahul Ramachandran; Kaylin Bugbee; Manil Maskey; Chris Lynnes. (2019). From ARDS to AODS: Future of analytics for Earth observations. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and Opportunities of Open Data in Ecology. *Science*, *331*(6018), 703–705. https://doi.org/10.1126/science.1197962

Robinson, N. H., Hamman, J., & Abernathey, R. (2020). Seven Principles for Effective Scientific Big-DataSystems. *ArXiv:1908.03356 [Cs]*. http://arxiv.org/abs/1908.03356

Roure, D. D., Goble, C., Bhagat, J., Cruickshank, D., Goderis, A., Michaelides, D., & Newman, D. (2008). myExperiment: Defining the Social Virtual Research Environment. *2008 IEEE Fourth International Conference on EScience*, 182–189. https://doi.org/10.1109/eScience.2008.86

Roy, D. P., Wulder, M. A., Loveland, T. R., C.E., W., Allen, R. G., Anderson, M. C., Helder, D., Irons, J. R., Johnson, D. M., Kennedy, R., Scambos, T. A., Schaaf, C. B., Schott, J. R., Sheng, Y., Vermote, E. F., Belward, A. S., Bindschadler, R., Cohen, W. B., Gao, F., … Zhu, Z. (2014). Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, *145*, 154–172. https://doi.org/10.1016/j.rse.2014.02.001

Safford, H. D., Sawyer, S. C., Kocher, S. D., Hiers, J. K., & Cross, M. (2017). Linking knowledge to action: The role of boundary spanners in translating ecology. *Frontiers in Ecology and the Environment*, *15*(10), 560–568. https://doi.org/10.1002/fee.1731

Satterfield, D. (n.d.). *Dan's Wild Wild Science Journal*. AGU Blogosphere. Retrieved October 9, 2020, from https://blogs.agu.org/wildwildscience/

Schöpfel, J., & Prost, H. (2016, November). Altmetrics and grey literature: Perspectives and challenges. *GL18 International Conference on Grey Literature*. https://hal.univ-lille.fr/hal-01405443

ScienceOpen. (n.d.). *ScienceOpen*. ScienceOpen. Retrieved October 9, 2020, from https://www.scienceopen.com/

Sentinel Hub. (n.d.). *Classification App*. Classification App. Retrieved October 9, 2020, from https://apps.sentinel-hub.com/classificationApp/#/

Sicular, S. (2016). *Three Architecture Styles for a Useful Data Lake* (No. G00303817; pp. 1–32). Gartner. https://www.gartner.com/doc/3380017/architecture-styles-useful-data-lake

SMD Science Management Council. (2018). *Science Mission Directorate Policy: Citizen Science* (SMD Policy Document SPD-33). NASA. https://science.nasa.gov/science-pink/s3fs-public/atoms/files/SPD 33 Citizen Science.pdf

Smithsonian Astrophysical Observatory. (2020). *Astrophysics Data System*. ADS. https://ui.adsabs.harvard.edu/

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115

Suber, P. (2012). What Is Open Access? In *Open Access*. The MIT Press. https://doi.org/10.7551/mitpress/9286.003.0003

The ENCODE Project Consortium. (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biology*, *9*(4), e1001046. https://doi.org/10.1371/journal.pbio.1001046

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. https://doi.org/10.1038/nature04226

Ton That, D. H., Fils, G., Yuan, Z., & Malik, T. (2017). Sciunits: Reusable Research Objects. *2017 IEEE 13th International Conference on E-Science (e-Science)*, 374–383. https://doi.org/10.1109/eScience.2017.51

Toronto International Data Release Workshop Authors. (2009). Prepublication data sharing. *Nature*, *461*(7261), 168–170. https://doi.org/10.1038/461168a

University of Cambridge. (n.d.). *Open Research*. Scholarly Communication. Retrieved October 9, 2020, from osc.cam.ac.uk

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428–436. https://doi.org/10.1016/j.jbusres.2017.12.043

Watson, J. (1990). The human genome project: Past, present, and future. *Science*, *248*(4951), 44–49. https://doi.org/10.1126/science.2181665

Woelfle, M., Olliaro, P., & Todd, M. H. (2011). Open science is a research accelerator. *Nature Chemistry*, *3*(10), 745–748. https://doi.org/10.1038/nchem.1149

Wulder, M. A., Masek, J. G., Cohen, W. B., Loveland, T. R., & Woodcock, C. E. (2012). Opening the archive: How free data has enabled the science and monitoring promise of Landsat. *Remote Sensing of Environment*, *122*, 2–10. https://doi.org/10.1016/j.rse.2012.01.010

Xu, C., & Yang, C. (2014). Introduction to big geospatial data research. *Annals of GIS*, *20*(4), 227–232. https://doi.org/10.1080/19475683.2014.938775

Yang, C., Yu, M., Li, Y., Hu, F., Jiang, Y., Liu, Q., Sha, D., Xu, M., & Gu, J. (2019). Big Earth data analytics: A survey. *Big Earth Data*, *3*(2), 83–107. https://doi.org/10.1080/20964471.2019.1611175

Zhu, Z., Wulder, M. A., Roy, D. P., Woodcock, C. E., Hansen, M. C., Radeloff, V. C., Healey, S. P., Schaaf, C., Hostert, P., Strobl, P., Pekel, J. F., Lymburner, L., Pahlevan, N., & Scambos, T. A. (2019). Benefits of the free and open Landsat data policy. *Remote Sensing of Environment*, *224*, 382–385. https://doi.org/10.1016/j.rse.2019.02.016

Zooniverse. (n.d.). *Welcome to the Zooniverse: People-powered research*. Zooniverse. Retrieved October 9, 2020, from https://www.zooniverse.org/

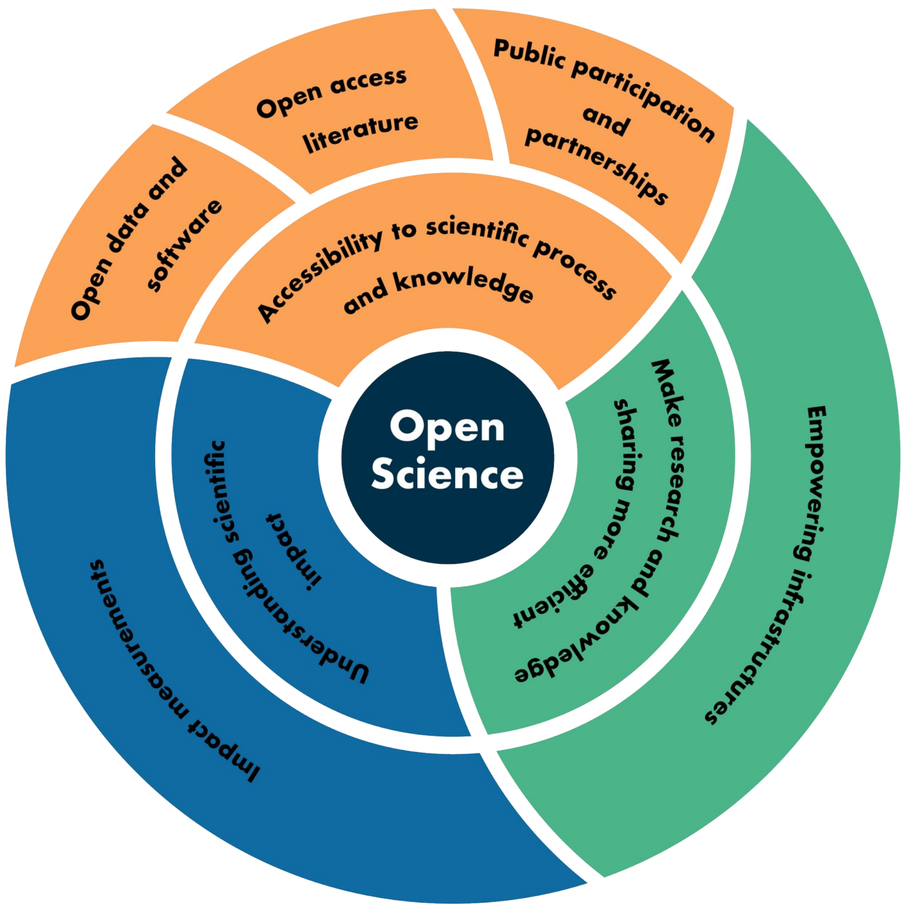

Figure 1: The open science concept represented as layers. The concept of open science is found in the center. The three open science focus areas are represented in the middle layer, while the outer layer represents data program-specific strategies that enable open science.

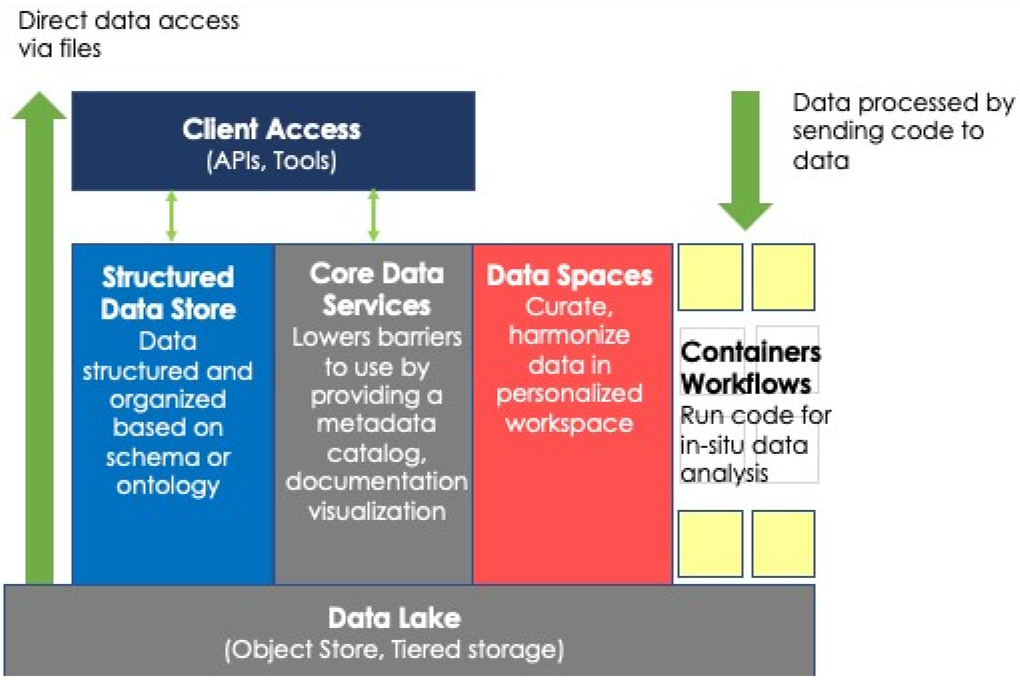

Figure 2.: A conceptual data system architecture in the cloud (Bugbee et al., 2020).