Table of Contents

1. Introduction

1.1. Goals of report

NOAA has a well defined Data Strategy [NOAA 2018] and Cloud Strategy [NOAA 2018] that clearly identifies goals, objectives and mechanisms to support open data, and NOAA has made substantial progress implementing these strategies. It is not our intention in this report to revisit or evaluate these strategies. Instead, our report has two goals.

Our first goal is to bring attention to selected issues around open data that were identified through surveys and interviews conducted by this NOAA SAB subcommittee (see Appendix A). Our second goal is to highlight certain issues related to NOAA's support for open science that were identified through the surveys and interviews. The NOAA Data Strategy and the NOAA Cloud Strategy are largely silent on open science (as distinct from open data).

1.2. Report's definition of open data and open science

The concepts of open data and open science are described in various ways. To facilitate the subcommittee work and the communication of our findings and recommendations, we utilized the following definitions for each concept.

1.2.1. Open Data

In this report, we use the definition of open data contained in NOAA's 2020 Data Strategy, which derives from the Evidence Act (II.a17):

> Open Data is a public data asset that is machine-readable; available (or could be made available) in an open format; not encumbered by restrictions, other than intellectual property rights, that would impede the use or reuse of such asset; and based on an underlying open standard that is maintained by a standards organization. [NOAA Data Strategy, July 2020].

Today, there is a general consensus on the definition of open data and many of the principles and processes for supporting open data. For example, open data is often linked with FAIR data principles, i.e. data that is findable, accessible, interoperable, and reusable (FAIR) as described in the 2016 publication "The FAIR Guiding Principles for scientific data management and stewardship" [Wilkinson 2016].

NOAA has made significant progress towards open data broadly following the goals and objectives listed in NOAA's 2020 Data Strategy, although as would be expected in an organization as large and as complex as NOAA, there is more work to do.

1.2.2. Open Science

There is less of a consensus on the definition of open science, although there is no lack of papers describing different visions of open science.

A 2018 literature review of 75 studies about open science identified several different definitions [Vicente-Saez 2018]:

- Open science as knowledge
- Open science as transparent knowledge
- Open science as accessible knowledge
- Open science as shared knowledge
- Open science as collaborative-develop knowledge

Overlapping with these definitions, some view open science as *reproducible knowledge* and focus on *open source software, collaborative environments for developing open source software, and software architectures* to support reproducible knowledge [NASEM 2018].

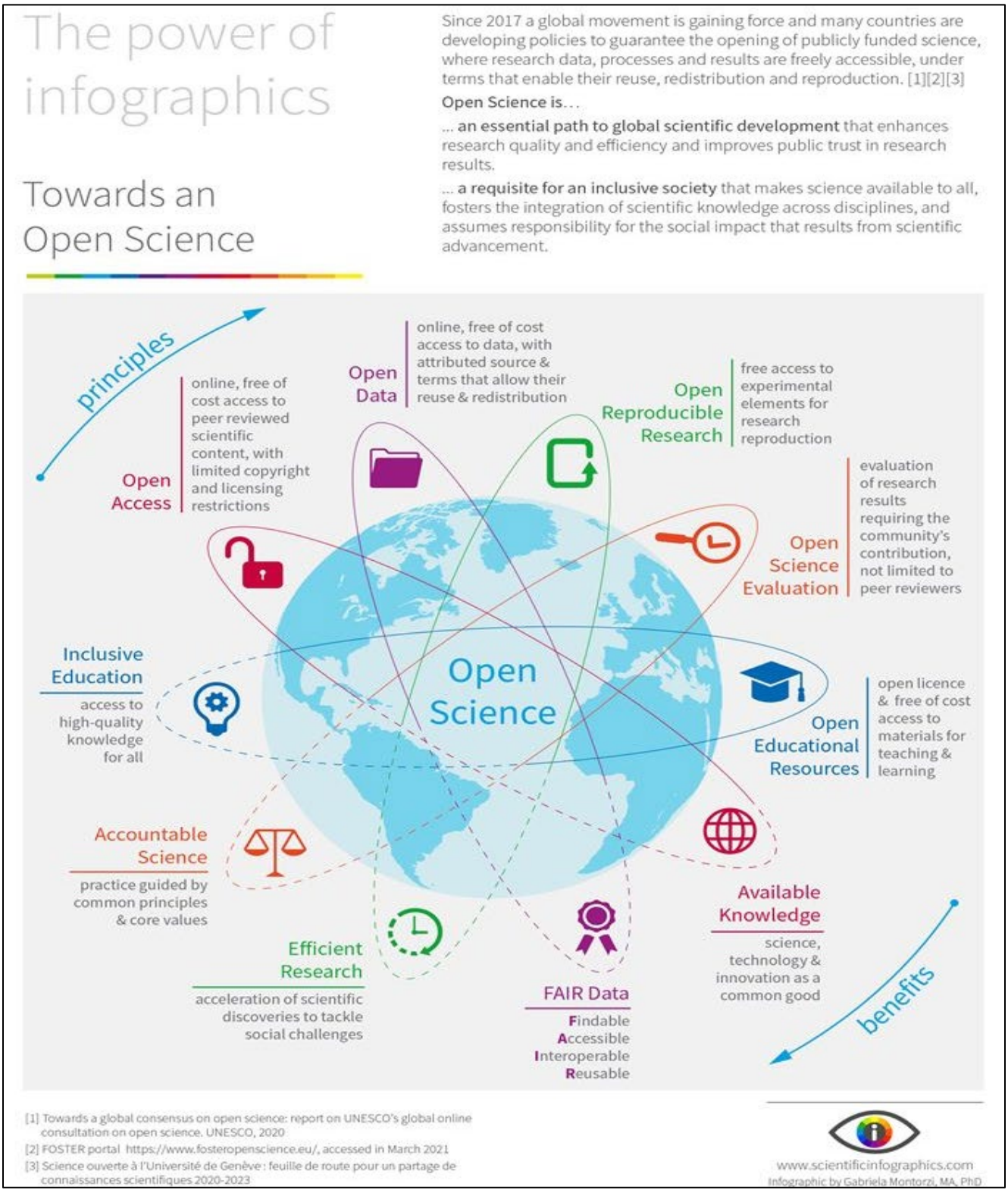

Figure: Open Science infographic showing the different components of open science. https://www.scientificinfographics.com/21-towards-an-open-science

In this report, rather than provide another definition of open science, we take the approach of stressing the importance of two open science guidelines:

1. In general, software developed by NOAA and NOAA supported projects should be open source with a permissive license that encourages engagement and reuse.
2. In general, research by NOAA and NOAA supported projects should be accessible and reproducible.

For definiteness, we use the definitions of open source software and reproducible research given by the National Academy of Sciences study on Open Source Software [NASEM 2018].

> Open source software. Software whose source code is under an open source license, by which the copyright holder grants to anyone the rights to inspect, modify, and distribute the source.
>
> Reproducible research. Research published with all the necessary data, source code, and configurations to run the analysis again, re-creating the results and data products.

To summarize, for the purposes of this report, we focus on two particularly important elements that contribute to open science: 1) the importance of open source software as a foundations for open science; and, 2) the importance of publishing research with necessary data, source code, configurations, and documentation required, in principle, to run the analysis again.

1.3. Context and Process

There are quite a few reports by different advisory committees, advocacy groups, and research communities around open data and open science. Many of these reports extend a year or more, gather material from numerous meetings, and produce substantial reports.

In contrast, our report collected 9 surveys from NOAA scientists and leaders (some surveys were filled out by two or more individuals). We had 7 virtual meetings where we held conversations between the committee and those that filled out the surveys. In the end, we gathered input from approximately 18 NOAA scientists, leaders, and outside experts (see Appendix B).

The goal of this approach was to survey the NOAA enterprise to gain insights into activities related to open data and open science and identify and highlight a small number of issues where additional focus may provide a disproportionate return in the support of open data and open science.

1.4. Outline of report

Section 2 describes some findings and recommendations from the surveys and our interviews with NOAA scientists, leaders, and others that we interviewed. Section 3 is a summary.

2. Findings and Recommendations

2.1 FAIR Data Across NOAA

The FAIR principles [Wilkinson 2016] are sometimes interpreted quite broadly, which increases the burden to implement them. Moreover, the FAIR principles can be challenging to implement for legacy datasets and products due to the lack of resources and the loss of relevant institutional memory. This is a problem given the many long-term archives of climate data from different sources including in situ, satellite, and model datasets that NOAA works with.

However, we emphasize that FAIR data is not absolute, the principles are aspirational. In addition, the originators of the concept note that a little FAIR-ness may go a long way towards achieving FAIR data [Mons, et al 2017]. This latter observation is reflected in our findings and recommendations.

We summarize this with a finding.

Finding. FAIR principles can be challenging to implement for legacy datasets and products due to the lack of resources and expertise.

The committee agreed that the core goals for FAIR, and thus open data, could be supported by consistently implementing the following four requirements for publicly available NOAA data across the enterprise.

1. All publically accessible NOAA data should have a persistent identifier (PID), associated PID-required metadata, and resources to keep relevant PID metadata up to date.
2. Publicly available NOAA data and its associated metadata should be, as much as possible, accessible through an open API.
3. All publicly accessible NOAA data should be associated with a generally accepted open data license agreement.

By a persistent identifier (PID) we mean a globally unique identifier associated with the data that is not associated with its physical storage system along with a system for dereferencing the identifier to find its specific storage location. In this way, the location of the data in a particular storage location can be changed with the PID remaining the same. PIDs are also known as Globally Unique Identifiers (GUIDs). Some datasets may have both PIDs and DOIs, and, in some cases, DOIs may be used instead of PIDs.

There is a broad consensus that best practices include the use of PID and of searchable metadata using in a common schema, such as CKAN and or the SpatioTemporal Asset Catalog (STAC).

Finding. Oftentimes, it is not always clear what minimum metadata is required for data to be FAIR. The cross-agency data enterprise could benefit from guidance on the minimum requirements within NOAA for making data FAIR.

NOAA Integrated Ocean Observing System (IOOS) Office guides regional associations' work by promulgating use of standards including those for open science and data. At this time, labs and organizational units in NOAA are addressing open science infrastructure and approaches independently. While NOAA has established a working group, there is no guidance on concepts or infrastructure at this time. This has resulted in inconsistent use of DOI across regional associations. Requiring data to be archived prior to issuing a DOI creates a bottleneck and single point of failure at the archives.

Finding: Without PIDs, finding data to reproduce or extend results is a significant barrier for scientists and reduces the scientific impact of NOAA's investments in data and data distribution.

The expansion of commercial production of science data is substantial and likely to expand both for space-based and in situ observations. These data are already forming a valuable contribution to NOAA's scientific enterprise but may create a barrier to reproducibility and scientific impact if not openly licensed. Additionally, there is additional management and risk enforcing license compliance. NOAA's position as a large federal partner and data distribution capabilities is valuable to these commercial partners.

Finding: The risk/reward of commercial data presents a challenge to NOAA's open data policies.

For example, cost considerations make it difficult to make data FAIR immediately at time of purchase. An approach suggested in the interviews, is to negotiate licenses so that after a suitable embargo period, data is made FAIR.

Recommendation 1. NOAA should follow the principles of FAIR open data and, whenever possible, these principles should be prioritized over other mission requirements.

1.1 Specifically, since FAIR is open to interpretation, NOAA data should all be required to have a PIDs, metadata, open-access APIs, and a standard open license (eg. cc-0 or cc-by).

1.2 Issuing PIDs for NOAA datasets should be made the highest priority and bottlenecks removed.

1.3 NOAA should consider the impact of any use restrictions on purchased commercial data on reproducibility and scientific impact and strive to minimize the use of non-open data whenever practical, as well as negotiating contracts that transition this data to open data after an appropriate time period.

2.2 Open Source Software

As mentioned above, there has been significant progress on open data at NOAA, but there has been less progress on supporting open science at NOAA through open source software.

From the surveys and the interviews, it was clear that ownership of software used by NOAA can be quite complex. Software code is often written over many years by many different organizations. When a NOAA partner or contractor writes code, it is copyrighted by that partner. For this reason, ownership of software code can be quite complex and it can be challenging to find primary authors and see if they agree on using an open source software license.

Finding. For projects that rely on software written over many years by different organizations, agreeing to license the software as open source software can be time consuming and challenging to achieve.

For new projects, an open science best practice is to agree to a common permissive open source software license at the start of the project that is used by all those involved and make a contractual term for any NOAA contractors working on the project and a clause in any cooperative agreements with NOAA partners.

Finding. Unless a common open source software license is used at the beginning of a project, it can be challenging later to make software open source.

Building a strong culture of open science at NOAA requires that software developed by NOAA projects be easily used by other projects, which requires appropriate open source software licenses. For this reason, in the surveys and interviews, it was recommended that existing and widely accepted open source software licenses be used by NOAA and that these be what are called “permissive” so that the software can be widely reused without undue restrictions.

Finding. Projects that create their own custom software licenses make it more difficult for other projects to reuse their software.

Open source software often relies on other open source software and open source software libraries and many of these are frequently updated, which requires that NOAA developed open source software be updated as well or that other steps be taken so that NOAA developed open source software continues to work as the open source software it relies on changes. A different but related challenge is that NOAA developed open source software must be updated as required as new security issues arise in the NOAA developed software and the open source software that it relies on. For these reasons, open software can not be written once and used for long periods of time, but must be kept up to date.

Finding. Keeping NOAA developed open source software working as the open source software it depends upon is updated and as new security issues arise requires occasional software maintenance, even if the underlying functionality doesn’t change.

Recommendation 2: Encourage and support the use of open source software as a key component of open science.

Recommendation 2.1 NOAA should develop agency wide guidance recommending the use of permissive open source software licenses for most projects, unless there are compelling reasons otherwise.

Recommendation 2.2 Any new projects should agree at the project start to use a widely accepted permissive open source software license and terms requiring this should be included in NOAA contracts and partnership agreements.

Recommendation 2.3 NOAA developed open source software will require maintenance over time and NOAA should develop agency wide guidance on how this will be supported if the project that develops the software is ended, while other NOAA projects that rely on the software continue.

2.3 Broadening the Engagement with the Open Science Community

Several interviews stressed a lack of training and understanding of what is meant by open science and open data and the mismatch between the current community requirements for open science and skills/training for the workforce can also be a limiting factor (in adoption of open science). A specific example was discussions of challenges related to the practical difficulties of releasing software in a community project and a need for guidance and clear directions for future software development and licensing.

Finding. Open science adoption requires additional workforce development.

Recall from Section 1.2 that for the purposes of this report, we focused on two aspects of open science: 1) open source software; and, 2) the importance of publishing research with necessary data, source code and configurations required to run the analysis again.

The importance of reproducible research in the context of 2) came up during the interviews. We summarize this with a finding.

Finding. An important component of open science is publishing research with necessary data, source code and configurations required to run the analysis again and obtain the same results and data products.

It is important to note several limitations around this finding:

- In many cases, the data itself is identified through PIDs, and not directly included in the published findings. This is especially important for large datasets.

- Software often ages very quickly, and virtual machines, software containers, and Jupyter Notebooks, and other commonly used mechanisms for rerunning scientific analysis often cannot be run even months later without updating various software libraries and modules. For this reason, the finding is scoped to encourage the release of the necessary software and configurations at the time the research is published. Of course, it would be nice to keep the software maintained and updated overtime, but this requires resources that are not always available.

NOAA has a long tradition of engaging with the scientific community and making its scientific research freely available. As the importance of publishing the software and configurations required to reproduce the results and data products becomes more widely implemented, it is important for NOAA to engage with the scientific community to establish best practices around this and other components of open science.

Finding: Best practices are still being worked out for how best to publish the software and configurations required to reproduce the results and data products of NOAA research.

Recommendation 3: NOAA should engage with the open science community around open reproducible research and support workforce training on how to do open, collaborative, and reproducible science in support of the NOAA mission. Specifically:

> **3.1 When NOAA scientists publish scientific papers, the software and configurations used for figures, tables, and core results should be made available at time of publication.**
>
> **3.2 NOAA should invest in workforce development in broad support of open source software, make research results reproducible at the time of publication, and more generally open science.**
>
> **3.2 NOAA should sponsor or leverage an annual conference or other annual event, such as a session at a larger scientific conference, with an accompanying report to engage consistently with the external scientific community around open data, reproducible research, and more generally open science.**

It is important to note that Recommendation 3.1 is scoped to providing software and configurations at the time of publication and encourages but acknowledges that often times there won't be the resources to maintain the software over time.

Finally, we note that an annual conference or other annual event around open science creates an opportunity for NOAA to be a visible leader who embraces open science that will hopefully make the agency more attractive to future employees.

2.4 Open Data and Open Science Across the NOAA Enterprise

A strength of NOAA is its commitment to open data and open science over many years, long before these terms became popular. As a consequence, the practice of open data and open science is distributed through NOAA line offices, with NOAA scientists and projects empowered to interpret open data and open science, which sometimes results in different interpretations and understandings of how best to support open science. At times, these differences can create some barriers to open science.

It is not unique to NOAA that policies and procedures developed at an agency level are not followed on an individual level due to different interpretations and implementation across NOAA of leading practices for open science. There is an inconsistent adoption of open science across regional associations and without any formal compliance assessment, this is likely to continue. NOAA's distributed nature is a source of strength, especially with regards to user communities where regional centers focus on understanding and serving regional needs. Yet, this distributed nature is also a source of duplicative efforts or different prioritization of enterprise needs. There is a strong need to apply open data and open science principles that are consistent across NOAA, NOAA's mission and benefit the community.

Finding. The distributed nature of NOAA data access points has created barriers to open science, some redundant workflows, and some inconsistent adoption of policies.

Recommendation 4: Consider providing consistent guidance across the agency for best practices, checklists, and dashboards to track adherence to open science principles, policies and mandates across the enterprise, while still supporting NOAA's distributed culture of data and science.

An entity such as the Data Governance Committee might be a candidate for these types of activities.

3. Summary

In this report, we have highlighted certain issues related to open data and open science that surfaced from surveys we sent out and online discussions we had with those that completed the surveys.

For this report, we focused on four aspects of making data FAIR (Section 2.1) and two aspects of open science. Specifically, we focused on the following two aspects of open science as a starting point: the importance of using open source software whenever possible (Section 2.2) and the importance of publishing or linking to the open source software, data and configurations required to reproduce research results and the associated data products at the time of publication (Section 2.3).

Moving to open data and open science is a cultural shift that needs visible leadership from NOAA and engagement and feedback from the NOAA research community, and the open

science community. All federal agencies are struggling with this change in the research enterprise. Resistance to change has at times hampered embracing new technologies and until NOAA addresses this with a clear mandate, backed up by promotion and support of those willing to embrace change, the move to open data and open science will be slower than the climate crisis urgently requires.

References

[Mons, et al 2017] 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/gfkrrv

[NASEM 2018] National Academies of Sciences, Engineering, and Medicine. Open Source Software Policy Options for NASA Earth and Space Sciences. National Academies Press, 2018.

[NOAA 2018] NOAA Cloud Strategy: Maximizing the Value of NOAA's Cloud Services, July, 2020, retrieved from https://sciencecouncil.noaa.gov/Portals/0/2020%20Cloud%20Strategy.pdf on November 1, 2022.

[NOAA 2018] NOAA Data Strategy, Maximizing the Value of NOAA's Data, July, 2020, retrieved from https://sciencecouncil.noaa.gov/Portals/0/2020%20Data%20Strategy.pdf on November 1, 2022.

[Vicente-Saez 2018] Vicente-Saez, Ruben, and Clara Martinez-Fuentes. "Open Science now: A systematic literature review for an integrated definition." Journal of business research 88 (2018): 428-436.

[Wilkinson 2016] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3, no. 1 (2016): 1-9.

Appendix A. Committee Members

Chelle Gentemann, Co-chair - NASA
Robert Grossman, Co-Chair - University of Chicago
Jason Hickey, Google (former Co-Chair)
Ilene Carpenter - Hewlett Packard Enterprise
Chris Lendardt - University of North Carolina, Chapel Hill
Zhaoxia Pu - University of Utah
Anthony Wu - AeroMarine LLC

Appendix B. List of Those Interviewed

Tom Augspurger - Microsoft
Mathew Biddle - NOAA, Integrated Ocean Observing System
Sid Boukabara - NOAA, NESDIS Office of Systems Architecture and Advanced Planning
Eugene Burger - NOAA, Pacific Marine Environmental Laboratory
David Fischman - NOAA, Office of Chief Information Officer
Kevin Garrett - NOAA, NESDIS Center for Satellite Applications and Research
Shane Glass - Google
Derek Hanson - NOAA, Office of General Council
Tony LaVoi - NOAA, Office of Chief Information Officer
Nancy Majower - NOAA, NMFS Office of Chief Information Officer
Joseph Pica - NOAA, National Center for Environmental Information
Ana Pinherio-Privette - Amazon Web Services
Dougals Rao - NOAA, National Center for Environmental Information
Karen Sender - NOAA, NMFS Office of Science and Technology
Adrienne Simonson - NOAA, Office of Chief Information Officer
Kim Valentine - NOAA, NOS Information Management Office
Tiffany Vance - NOAA, Integrated Ocean Observing System
Melissa Zweng - NOAA Integrated Ocean Observing System

In addition, other NOAA scientists and leaders answered some of the questions in our surveys, but did not participate in our virtual meetings, and, for that reason, are not listed here.