TOPS Open Science 101

Table of contents

Transform to Open Science

Welcome to the open science guide for Transform to Open Science (TOPS).

What is Open Science?

The United States government defines open science as "the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility, and equity."

They believe that open science—-opening up the scientific process from idea inception to result—-increases access to knowledge and expands opportunities for participation. Sharing the data, code, results, and knowledge associated with the scientific process enables more inclusive, diverse and equitable participation in science, while also leading to more transparent, replicable, and reproducible results. But achieving this openness requires changing how we work, to help us move forward together.

The Open Science Journey

Research labs, scientific funding organizations, and individual researchers have known and discussed for many years how interdisciplinary and diverse teams are capable of advancing scientific progress. These groups and individuals began to advocate for inclusive labs and organizations; places where data and scientific practice was equitable and accessible to people from different backgrounds, with differing levels of academic training, and with different lived experiences. Although they may not have called this movement towards diverse and accessible research "open science," these same principles of equity and inclusivity are core to the open science ethos.

Other researchers and organizations have come to advocate for open science through their experiences trying to access data, code, research methods, and publications through the course of their own scientific practice or funding apparatus. Frustration with embargo periods, incomplete or unsorted data sets, non-replicable results, or code that is anything but user-friendly have all resulted in a movement for full transparency of research, from the idea inception through the pre-registration of studies to the final results via open-source code, public datasets, and open-access publications.

This guide is for you, your team, or your organization to become more involved with this movement. *We are so glad that you are here on the road to open science with us!*

1 Open Science 101 Curriculum

1.1 Curriculum Overview

Transform to Open Science (TOPS) is an initiative designed to rapidly transform agencies, organizations, and communities to an inclusive culture of open science. Developed by the TOPS initiative, the guidance provided by the Open Science 101 curriculum will promote the transformation of the research landscape and the accompanying advancement of scientific discovery. The Open Science 101 curriculum aims to introduce learners to a nuanced understanding of open science, enabling participants to better understand an open science workflow from end to end. The focus of the curriculum will be on providing learners with a basic understanding of open science, its ethos and benefits, and how to actively participate in open science communities.

1.2 Open Science Curriculum: Open Science 101

1.2.1 5 Modules Organized as a Scientific Workflow

Part I

OS101 Module 1: The Ethos of Open Science

Welcome to The Ethos of Open Science

About This Module

Welcome to this introductory module on open science. Open Science is the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility, and equity. In this module, you take a closer look at what open science is, the current landscape as well as the benefits and challenges. You then get a glimpse into the practice of open science including a case study. To start your journey with open science, you are presented with actions that you can take starting today, such as exploring communities that you can engage with.

Target Audience

This module is for anyone who is interested in open science and would like to learn the benefits and ways to get started today.

Learning Objectives

After completing this module, you should be able to:

- Explain what open science is, why it's a good thing to do, and list some of the benefits and challenges of open science adoption.
- Describe the practice of open science, including considerations when writing a management plan and the tasks in the "Use, Make, Share" framework.
- Evaluate available options when determining whether research products should or should not be open.
- List ways to connect with others who are part of the open science community.

Key terms

Select the term to see the description.

Open Science – The principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility and equity.

Open Data – Data that can be freely used, re-used and redistributed by anyone. Learn more here.

Open Source – Computer programs in which the source code is available to the general public for use or modification from its original design. Learn more here.

Open Access – Free access to information and unrestricted use of electronic resources for everyone. Learn more here.

Interdisciplinary – Combining multiple academic disciplines into one activity, such as a research project. Learn more here.

Equitable – Indicates the absence of unfair, avoidable or remediable differences among groups of people, whether those groups are defined socially, economically, demographically, or geographically or by other dimensions of inequality (e.g. sex, gender, ethnicity, disability, or sexual orientation). Learn more here.

Citizen Science or Community Science – The practice of public participation and collaboration in scientific research to increase scientific knowledge. Learn more here.

Open Research – How research is performed and how knowledge is shared based on the principle that research should be as open as possible. Learn more here.

Open Scholarship – An expansive term meant to encompass the rapid and widespread sharing of a range of scholarly activities and outputs across multiple disciplines. Learn more here.

Reproducibility and Replicability – Reproducibility is defined as obtaining consistent results using the same data and code as an original study (synonymous with computational reproducibility). Replicability means obtaining consistent results across studies aimed at answering the same scientific question using new data or other new computational methods. Learn more here.

Peer Review – The evaluation of work by one or more people with similar competencies as the producers of the work - that is, the authors' peers. Learn more here.

FAIR principles – Principles to improve the Findability, Accessibility, Interoperability, and Reuse of digital assets. The principles emphasize machine- actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

Metrics (in context of scientific merit) – Quantitative tools used to help assess the quality and impact of research outputs (eg. scientific articles, researchers, and more). Learn more here and here.

Altmetrics – Alternative tools to assess the impact of a scientific article that do not involve journal-level usage information like impact factors. Learn more here.

Openness – A concept or personality trait that involves transparency, collaboration, honesty, and receptivity to new ideas and experiences.

Transparency – The quality of being easy to perceive or detect.

Rigor – Widely used by educators to describe instruction, schoolwork, learning experience, and educational expectations that are academically, intellectually, and personally challenging.

Computational Provenance - Seeks to develop systematic, computationally-based processes and standards for capturing, and making available for us, information about who created an object, when it was created or modified, and the process or procedure that modified the object.

Navigation

Lesson 1: What is Open Science?

Lesson 2: Why is Open Science Important?

Lesson 3: How to do Open Science

Lesson 4: When Not to be Open

- Common Fears Around Openness
- Misaligned Incentives
- Social Barriers
- Institutional and Infrastructure Barriers
- Lesson 4: Summary
- Lesson 4: Knowledge Check

Lesson 5: Planning for Open Science: From Theory to Practice

- Planning for Open Science
- Designing for Openness
- Case Study: The Outcomes of an Open Plan
- Steps to Continue Your Open Science Journey
- Lesson 5: Summary
- Lesson 5: Knowledge Check
- The Ethos of the Open Science Summary

2 Lesson 1: What is Open Science?

2.1 Navigation

2.2 Overview

In this lesson, you take a closer look at what open science means, including the intended goals and outcomes of adopting open science as an individual and as part of a larger community. You then review examples of open science in action. Finally, you wrap up the lesson by taking a closer look at why adopting open science is needed.

2.3 Learning Objectives

After completing this lesson, you should be able to:

- Explain the motivation to do open science and the goals of open science.
- Define open science.
- List different groups that practice open science.

2.4 Motivation for Open Science

Welcome to the first module of the TOPS Open Science 101 curriculum! Module 1 has five lessons that present information about the Ethos of Open Science. This incorporates the motivations and best practices for making science more open. This course was made possible thanks to the work of our Transform to Open Science (TOPS) team, open science Subject

Matter Experts (SMEs), and the entire TOPS community (3100+)! Please note that all image attributions are given at the end of each module.

We are really glad you are here!

This is the first lesson in the module on the Ethos of Open Science. Let's begin by explaining the word "ethos".

"Ethos is the distinguishing character, sentiment, moral nature, or guiding beliefs of a person, group,"

Merriam Webster

This lesson describes and showcases what makes Open Science, as an approach to knowledge production, unique or distinguishable from other scientific approaches.

Note that "ethos" is not exactly "ethics", but offers a broad enough term to include the moral attitudes held by the individuals or institutions who practice open science. To clarify the moral element to this discussion, we speak of "responsible open science" going forward.

The lesson introduces the concept of open science as a whole by explaining its motivation, definition, and operation. The lesson then reviews different components of science and the pillars that make them up. Throughout this module, we have integrated ethics around open science that dictate how you share, give due credit, and work together. "Practice the Golden Rule" - treat others the way you would like to be treated in their situation.

The curriculum of this course builds upon the work of a broad community from across the globe and fields of research who paved the way to open science.

2.4.1 The Internet Facilitates Sharing of Information

Historically, factors like time, access to sufficient tools and data, and physical proximity limited who could be involved in science, as well as how easily collaboration could take place within the scientific community. More recently, digital resources like the Internet have increased participation by eliminating barriers to entry and presenting a platform for digital collaboration on a global scale. The internet offered people access to the appropriate infrastructure to conduct open science, while the practices of open science enabled more people to engage with research products. Unfortunately, challenges remain for people who don't have the right computational tools and/or speak the relevant languages.

The Internet creates many outlets for public hosting and free access to research and data. These outlets combined with advances in computational power enable nearly anyone to perform complex data analysis. It is now possible to connect participants, stakeholders, and outputs of

open science on the Internet to make scientific processes and products easier to discover and access.

2.4.2 Why Should We Do Open Science Now?

Science and science communication increasingly face severe pushback from the public because of inadequacies in the reproducibility of results and the spread of misinformation, respectively, that foster mistrust. The practice of open science counteracts this by involving community feedback to validate results in a more robust manner and combats misinformation by making results available to the public.

Reproduce Results

Science becomes more robust and accurate when scientists validate their colleagues' results. However, the rapidly-growing pool of published research presents an overwhelming challenge to reproduce:

Source: https://www.pnas.org/doi/full/10.1073/pnas.1708290115

- In 2011, the AAAS, publisher of Science, began requiring the authors of computational research reports to share data and software upon request
- In 2018, a research study was carried out that investigated 204 articles for reproducibility and that were published in the journal Science after 2011. It was found that only 26% of papers were able to be reproduced, with the two primary reasons being the inability to get access to the data and software and the fact that the methods were not described in sufficient detail.

How many studies were reproduced?

No Yes

Open Results Enable Iteration and Improve Error-Detection

In this section, we will look at an example of how closed science can restrict research impact by following the outcome of a highly cited journal article to understand how science functions to inform a field's state of research, the decisions of policy makers, and the actions of society.

A 1990 analysis of satellite data on climate temperature concluded that the upper atmosphere experienced no warming, a finding that contradicted early climate models predictions. Policy-makers concluded from this result that researchers don't understand climate models enough to warrant changes in environmental policy. The processed data from this study were made open-access but, as was typical for the time, neither the original data nor the code used for processing and analyzing the data were shared by the original research team. Eight years after the article was published, other scientists noticed that the original authors didn't account for several important effects. This oversight introduced errors into the dataset and falsely produced artificial cooling to the temperature measurements. It took another five years and

additional funding to reproduce the code and conduct a new analysis. Thirteen years after the original paper, it was confirmed that the upper atmosphere was warming and agreed with climate model predictions.

The inability for the scientific community to access an article's original data and code slows the pace of discovery, thirteen years in this case, and forces other research teams to repeat the work (code) instead of moving on to new projects. This isn't the pace that we want to advance science, with one step forward and two steps back to iterate and resolve problems.

The intentions of the original research group were not to conceal or prevent access to their data and methods; the community norms at the time simply did not include the sharing of data and software openly. This is, in part, because it allows researchers to keep a competitive advantage when seeking funding opportunities. In this case, the research group simply followed this common practice. This culture of closed science needs to be changed because the practice of withholding code (or data or other research artifacts) can stifle scientific progress. In the climate change example, a flawed study could have swiftly been corrected by open peer feedback but it instead undermined the credibility of climate scientists. The cost to progress on climate change research and the prevented benefit to society was enormous. It is imperative to shift the entire science ecosystem, policies, and rewards towards the prioritization of openness if the full and immediate benefits of research are to be realized.

Traditional Publishing Limits Participation

Historically, scientific publishers have charged subscription fees to access journals and, often, article processing charges (APCs) to cover the costs of preparing a manuscript for press (even when the peer reviewers were volunteering their time). These practices limit both who could read papers and who could publish results.

Open access publishing has significantly increased the number of articles that are available as electronic copies online. A growing number of governments and funding agencies are starting to mandate that research funded by taxpayers must be accessible to the public after publication. However, the current hybrid system still does a poor job of allocating costs fairly across the research publication process (more on this in Module 5).

The issue of who has access to published papers also motivates open science. For example, even though more climate research is made available as open access than that from other scientific fields, the majority of climate research articles, including many important ones, remain behind paywalls. Climate misinformation is freely available to anyone online but scientific climate results are mostly hidden from the public behind paywalls. This practice does not increase trust in science.

Source: https://www.unesco.org/en/articles/can-science-be-more-equitable-so-everyone-enjoys-benefits-open-science-answer

2.5 What is Open Science?

But what is open science exactly? To illustrate, first we'll present a definition for open science that was developed in 2023 by the U.S. federal government.

"Open Science is the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility, and equity."

The White House Office of Science and Technology Policy Memo, 2022 (adapted)

This definition was developed by a team of U.S. federal agencies. They reviewed many different definitions of open science from both within the U.S. and around the world. These definitions have changed over time as the understanding of open science has matured and become more nuanced. Let's break down the definition a bit more:

- Research products and processes should be available to all, not just a small subset of experts, particularly if funded with public funds.
- Research products and processes should be 'respecting diverse cultures' – fostering an open dialogue between researchers, indigenous people, and local communities. This also means that research must respect the diversity of laws and customs in different countries and/or as they apply to different kinds of research.
- While open science is our aim, security and privacy remain important concerns. Therefore, select sensitive information should be protected.
- Of the stated principles, "Fostering collaborations, reproducibility, and equity", the first two are research standards, while the latter refers to the inclusion of people who might otherwise get left out.

Open science is a culture intended to promote science and its social impact. Open science creates new opportunities for different stakeholders including researchers, political decision makers, and public participants. Open science increases study transparency, repeatability, reproducibility, and confirmation. We expand what these terms mean and why they matter throughout this module and later in Open Science 101 modules.

2.5.1 How Do You Do Open Science?

The Ethos of Open Science is a broad term that encompasses the moral and ethical attitudes held by individuals and institutions about practicing 'open' science. There is an ethical element to sharing both new knowledge and the processes used to obtain said knowledge. It is important to note that there is no one be-all way of practicing or conducting open science.

Diverse practices, assumptions, and goals are just part of the complexity of open science. There are also divergent moral principles that guide open science communities. Such principles are captured in "codes of conduct". A code of conduct is a community governance mechanism that outlines the principles and practices expected of a given research community's members, as well as the process for investigating and reprimanding those in violation of the code.

In a sense, a code of conduct constitutes the moral backbone of a research community. However, as with the numerous schools of thought, there are similarly many codes of conduct. In other words, there is no one set of universal principles that all open science practitioners abide by. For example, consider how OLS, INOSC, allea, AGU and Ethical Source all have different codes of conducts and guiding principles.

This great diversity responds to the growing proliferation of open science initiatives and the great use we can make of open science approaches to knowledge.

The goal of TOPS is to push the community as a whole towards open science as an ethical responsibility to share knowledge.

Ethics	Morals
Community consensus about "proper" behavior.	Personal set of standards for "good" behavior.

You may or may not have heard the term "Open Science" used before this course. If you have, you likely have some preconception of open science and what it looks like in practice. Let's collectively look at those various conceptions.

2.5.2 Activity 1.1: Think About the What and How of Open Science

In this activity, reflect on your answers to the questions and then compare your thoughts to the key takeaways.

- What does the act of open science look like? Does a scientist use or create something specific that would characterize their research as open? What comes to your mind?
- Describe how you currently share your materials (data, code, results)?
- How might you share materials in the future more openly?
- What stands in the way?

Key Takeaways:

There are many ways to do open science. One of the best ways to start your open science journey is to think about how you do your science and how you might be able to make your science more open.

Goals of Open Science

The nuances of implementing open science take time and experience to develop. This curriculum provides practical steps and insightful examples to make open science happen.

Goals of open science include: - Facilitating collective benefit. - Achieving inclusive development and innovation. - Realizing equitable outcomes. - Ensuring appropriate control over the data you use, make, or share throughout your research process, and comply with policy, regulatory, and legal guidance on release. - Nurturing respectful relationships with the culturally-diverse communities who might provide or interact with your data, software, and results. - Actively engaging with interested communities to increase representation and consultation throughout your research efforts. Members of any communities that provide data must be allowed to determine the benefits, harms, and potential future uses based on their community values and priorities.

2.5.3 Fostering Collaborations, Reproducibility, and Equity

The IDEA of Open Science: Inclusive, Diverse, Equitable, and Accessible.

Openly using, making, and sharing research analyses, software, or datasets gives everyone credit for their work.

Sharing is grounded in the belief that access to information and the ability to collaborate is essential for advancing scientific understanding and solving complex problems.

Open sharing enables greater transparency in the scientific process and facilitates reproducibility; it enables collaboration and inclusion of more diverse perspectives and expertise; and it makes scientific knowledge more accessible to the public.

Not only does open sharing help society, but it also can benefit each of us as individual researchers. It can lead to greater visibility, impact, and credit of your results, data, and software; it can provide access to new collaborations and ideas, and it can fulfill ethical and social responsibilities.

2.5.4 Activity 1.2: Definition of Open Science

2.5.4.1 Match the parts of the OSTP definition with the relevant IDEA/Respecting cultural diversity concepts.

Engage with interested communities to increase research efforts

Respecting cultural diversity representation throughout

Fosters inclusivity, diversity, equity and accessibility

IDEA

2.5.4.2 Fill in the missing word in the definition of open science.

"Open Science is the principle and practice of making research products and processes _____ to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility, and equity."

- The White House Office of Science and Technology Policy (OSTP) Policy Memo, 2022

2.5.4.3 According to the White House Office of Science and Technology Policy, open science fosters...

Select all that apply.

- Collaboration
- Competition
- Equity
- Limited accessibility
- Reproducibility

2.5.5 Examples of Open Science in Action

EXAMPLE 1 : CITIZEN SCIENTISTS DRIVE NEW RESEARCH METHODS

EXAMPLE 2 : RADAR DATA TO KEEP AN EYE ON CLIMATE CHANGE

JunoCam is a camera/telescope aboard the Juno spacecraft that orbits Jupiter. It was added to Juno specifically for citizen scientists and public outreach purposes.

The camera's original purpose wasn't to perform science, only observe. However, access was given to citizen scientists who processed pictures by stitching JunoCam image bands together. Participation among the citizen science community helped expand the utility of the camera and led to an unforeseen success. Iconic images created by these citizen scientists of Jupiter and Europa have intrigued and inspired people from across the globe.

Collected data also benefited traditional science:

"Pretty much all of what we are learning about the structures and dynamics of Jupiter's clouds is coming from publicly-edited images," says planetary scientist and JunoCam wrangler Candice Hansen. Hansen explains, "The team is processing a few images itself but with no image processing staff, the researchers are relying on the work of citizen scientists."

The Juno spacecraft team made new and unexpected science possible through open and easily accessible data. Public participation brought new perspectives and expanded the team's capacity to conduct valuable research.

Credits:

Image data: NASA/JPL-Caltech/SwRI/MSSS

Image processing: Navaneeth Krishnan S © CC BY

EXAMPLE 1 : CITIZEN SCIENTISTS DRIVE NEW RESEARCH METHODS

EXAMPLE 2 : RADAR DATA TO KEEP AN EYE ON CLIMATE CHANGE

Have you ever seen weather forecast images for your location? That data comes from NEXRAD radar stations, many of which have been operating for over 30 years. The data has always been made publicly available, but can be difficult to use. It was mostly used for rain information, so stations didn't see a need to make it readily accessible after 24 hours. Users who wanted the historical data from NEXRAD had to work through the following arduous steps:

Go to a website.

Make a request (but not one too large).

Wait for a robot to read the data off tape storage and copy it online.

Receive an email with instructions on where to download a user's data.

Download the data.

The massive size of the dataset, more than 250TB, made it essentially impossible to do large-scale analysis. Nobody had the time to make these requests and download the data bit by bit.

However, in 2015, all NEXRAD data were moved to and made freely available in the cloud. Usage of the dataset increased almost immediately!

Researchers started using the NEXRAD data for other types of science. For example, they used NEXRAD radar readings of birds to monitor flight patterns. In particular, purple martins! Purple martins form huge roosts of up to 50,000 birds that can be tracked using radar. The purple martins perform stunning aerial performances that can now be tracked with the same technology previously reserved for rain measurements.

In another example of new NEXRAD uses, a NASA-led study linked variability in bird migration to large-scale climate patterns that originate thousands of miles away. The better land managers understand current migration patterns and foresee behavioral changes in these birds due to climate change, the better they can direct their conservation and habitat restoration efforts. The newly- accessible radar data provides valuable insight needed to achieve their goals. This study was funded by NASA, uses NOAA NEXRAD data, and made fully available for the first time by the AWS Public data program.

The science of the future makes data so easy to use that it allows new questions to be asked, and answered, quickly to benefit society and policy.

2.5.6 Key Takeaways: Definition of Open Science

Open science is multi-faceted and can mean different things to different people. Some scientists may think of open science as open access publications or as citizen science. In this module and more broadly in this curriculum, we make it clear that open science encompasses all of these various aspects. Right now, you may be more comfortable with some aspects of open science than others - and that's ok! We hope that you will better understand the scope and nuances of open science by the end of this curriculum, and also understand the benefits of doing open science, as we have begun to discuss in this section. Specifically, the increasing amount of information about scientific processes and products available online enables access for a much wider array of individuals and communities to use, make, and share scientific research and results.

2.6 Who Does Open Science?

As briefly discussed in previous lessons, open science doesn't only involve researchers; many other stakeholders are affected by the outcomes of open science. Stakeholders include any individuals who can affect or be affected by open science projects.

Scientific research should benefit humanity. Although open science has many stakeholders, the advantageous interaction between science and society takes place among three core groups: scientific researchers, policymakers, and the public. Researchers do science and share their results with policy makers and the general public to inform their decisions and improve their lives. The public helps to fund research through taxes and can provide input to future areas of study. Policymakers help to implement measures that are informed by scientific results to improve the health, environment, and livability of society.

These three stakeholder groups remain central to the world of open science. However, the inclusive nature of open science demands participation from the broader public. Growth in public participation in science can occur by removing barriers to those historically excluded and by expanding the community of people who support the scientific research itself.

Here we list some core groups who we envision as taking part in and/or benefitting from open science, while being fully aware that this list is not exhaustive and the categories we choose here have very blurred boundaries.

Select each tab to get more information.

RESEARCHERS

POLICY-MAKERS

GENERAL PUBLIC

Researchers are often thought of as the ones who do open science to benefit others. However, researchers themselves can also greatly benefit from open science. Their work can achieve higher visibility among colleagues and the public, they receive credit for a full range of activities related to their science (including time spent sharing data and code, for instance), and they have more access to datasets.

A team of supporters and collaborators enables this research to take place. Open science aims to include these supporting members of the scientific process and ensure they receive credit for their contribution to improve science.

RESEARCHERS

POLICY-MAKERS

GENERAL PUBLIC

Policymakers represent another key community in the science environment. Policy makers can reference scientific findings to inform their decisions for the betterment of society. Those who help in the understanding and dissemination of these policies (including educators and science journalists) are crucial to this process. Policy makers can also play important roles in ensuring and facilitating open science by setting data management processes, encouraging open access legislation, and developing ethical guidelines for experiments. Policy makers can benefit from open science by gaining better access to scientific output via the open sharing of research results.

RESEARCHERS

POLICY-MAKERS

GENERAL PUBLIC

The public plays a crucial role in science today as consumers of scientific results who make decisions based on, and adhere to policies shaped by, scientific results. Open science can make scientific results, data, and workflows more accessible to the public by strengthening routes of access to trustworthy sources of information, which in turn increases trust in science. The public can also take part in open science through community science projects, for example as volunteers to collect or manage data. As a result, participants boost their understanding of science and feel empowered through opportunities to exert influence.

Open science can strengthen the connection between all of these groups. Communication between researchers and both the public and policymakers stands to drastically improve with more transparent and accessible scientific knowledge.

2.6.1 Activity 1.2: Think about Open Science

In this activity reflect on your answers to the questions and then compare your thoughts to the key takeaways.

- In your field, what steps are being taken to increase openness, and what stands in the way?
- What could help to increase openness?
- What stands in the way?

Key Takeaways: How Closed Science Affects You

Some observational reports are difficult to replicate because of the uniqueness of the observations or ethical considerations (for example, unique astronomical events and long-term ecological observations, or accidental experiments like crashes and inadvertent releases of contaminants).

Here, we would like you to consider limitations brought about by software, data, or methods that are not fully described or made accessible.

2.7 Lesson 1: Summary

In this lesson, you learned:

- The motivation for open science as well as its goals and outcomes.
- Why we should be doing open science now and how technology has made it more achievable.
- The definition of open science.
- Different groups that do open science.

2.8 Lesson 1: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/04

What is open science?

- Sharing all information so anyone can access it.
- A legal definition of free science.
- A nuanced way of sharing the process and products of science.

Question

02/04

Why is open science happening now?

- Communicating research results is an integral aspect of science.
- The internet and increasing availability of computers enables access by a much wider segment of society.
- Researchers are mandated by law to share all their findings and raw data immediately after conducting their experiments.
- Open Science is happening now because of a global shortage in academic professionals, necessitating widespread public involvement.

Question

03/04

Select all open science actions that foster collaboration, reproducibility, and equity.

- Open sharing of data, software, and results.
- Prioritizing the fastest possible publication speed to ensure that results are disseminated quickly.
- Ensuring that the most resources are allocated to the most popular and widely reported fields of research.

Question

04/04

What societal problems can open science help to address? Select all that apply.

- Scientific reproducibility crisis.
- Spread of misinformation.
- The increase in paper usage due to the printing of scientific journals and articles.
- The overpopulation of certain animal species used in laboratory testing.
- The issue of decreasing interest in artistic and cultural studies due to the emphasis on scientific research.
- The rising costs of luxury consumer goods are influenced by technological advancements.

3 Lesson 2: Why is Open Science Important?

3.1 Navigation

3.2 Overview

In this lesson, you learn how adopting open science benefits you as a researcher and society. You also learn about some of the challenges and hurdles with using open science principles and how to navigate them.

3.3 Learning Objectives

After completing this lesson, you should be able to:

- Describe the ways in which open science benefits your career with attribution, reach, and more collaborations.
- Describe the ways in which open science can advance science.
- List the benefits society receives when open science principles are adopted.

3.4 Open Science Breaks Down Stovepipes and Increases Innovation

"We need more WE science rather than ME science."

Harlan Krumholz,

The world faces both known and unforeseen challenges. These dynamic challenges will demand a new approach to science that achieves success through a responsive and inclusive scientific ecosystem. This requires more diverse teams - teams with more hands, eyes, and brains that have diverse experiences to participate.

In the previous lesson, we learned about foundational concepts that define open science – its importance, its purpose, and its proven successes. In this lesson, we address key benefits and challenges of implementing open science principles in research activities.

This lesson presents a perspective on the development of science that focuses on the intersection of scientific results, the process of creation, and the stakeholders that constitute the community.

This lesson highlights the benefits of open science to a wide range of stakeholder groups, along with examples that invite readers to investigate further. Additionally, this lesson explores common challenges to adopting open science practices.

Figure: There are many benefits of open science. CC-BY Danny Kingsley & Sarah Brown.

3.5 Benefits to You

3.5.1 You are Your Best Future Collaborator!

Doing open science not only lets other people understand and reproduce your results, but lets you do so as well! Implementing open science principles such as good documentation and version control helps you, potential collaborators, and anyone else to understand your results.

If your work is shared publicly, you will never lose access even if you move institutes or change jobs. Many researchers move around institutions and organizations. By having your data and software and results in repositories, you will always have access to them.

Implementing best practices for open science in your work not only helps you document, it could strengthen your funding proposals. Funding agencies have begun to realize that openly sharing research products can increase their citations received and uptake, resulting in a better return on investment.

Well-documented research products also demonstrate the quality of your work, which helps with public communication efforts and can also attract better collaborators. Reliability and a strong work ethic motivate others to want to work with you.

3.5.2 Give and Get Credit When Using Results of Others

In addition to documenting your own research, the practice of giving credit to everyone who has contributed will strengthen your scientific community reputation and actualize the shared values of open science. As people gain confidence in the benefits of cooperative research, they will also start giving credit to more contributions that might previously have gone unacknowledged. Different work performed as part of a paper can be given in an author contribution statement like the example shared here.

The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 license. DOI: 10.5281/zenodo.3332807.

3.5.3 More Visibility and Impact

In addition to improved scientific accuracy, adhering to open science practices potentially offers personal career benefits to researchers. Openly published research has significantly more visibility and impact potential with large audiences across the internet, which can lead to more citations, like-minded collaborators, and career/funding opportunities, according to a 2016 study.

3.5.3.1 Emerging evidence that some aspects of open science can increase your citations.

Publishing open access increases citation count by 18%, according to a 2018 study.

Articles that make their data openly accessible via a direct link to a repository see ~25% higher citation impact, according to a 2020 study.

Publishing as open access may have prohibitive costs for some researchers depending on the venue. There are often other options that allow authors to share their work freely and openly. In Module 5 on Open Results, we discuss some of these other options including preprints and diamond open access.

There are many different research outputs that can be openly shared and made citable:

- Code
- Data
- Research talk slides
- Lectures
- Blog posts
- And more!

All of these are tangible, scientific outputs! Much of our time as researchers is spent writing code, collecting data, putting together lectures and not *just* publications. Publicly sharing materials publicly makes receiving a citation more likely.

3.5.4 More Collaborations

Open science practices can also enable stronger collaborations, both within and between disciplines, as evidenced by a 2016 study. The ease of access to open data brings new agents to the landscape that allow for broader and more diverse participation. Through open science practices, such as pre-registration where researchers document their research plan at the start of a study, one allows for a stronger research design because feedback from various collaborators and stakeholders can be solicited before data collection begins. Similarly, preprints allow for swifter feedback on conclusions drawn from data once it is collected.

3.5.5 Activity 2.1: Benefits to You

In this activity, reflect on your answers to the questions and then compare your thoughts to the key takeaways.

- Can you find your own previous work, post-publication and/or pre-publication? Can you bring your research materials (data, code, results) with you if you change institutions?
- Can you find the work of your collaborators? Of scientists in other fields that you find interesting? Have you reached out to others to collaborate with them after finding interesting results?
- Are people in your field giving and getting credit for work done?

Key Takeaways: Benefits to You

- Being more open encourages best research practices and makes it easier for you to build on your work.
- Open results have more visibility and impact.
- Open science encourages more collaborative science.

3.6 Benefits to Science

3.6.1 Transparent Science is Reproducible Science

When computers are used to produce scientific research, the code is considered a "method". Much like a lab research setting, a set of instructions for working with cells or agar plates can be considered a method. Peer-reviewed methods are an essential step in the scientific process. When these steps are not shared, no-one else can reproduce the work or build upon it for

future scientific endeavors. Open methods allow people to judge whether or not the methods are trustworthy. In Lesson 1, the story of the Global Cooling Error presented a poignant example of science that was not reproducible because of a lack of data transparency.

3.6.2 Open Science Can Improve Accuracy

A study from 2022 found that researchers who practice transparency and promote verifiability benefit from readers and stakeholders who judge whether results presented are accurate and, according to a related study, that the results are not produced by questionable research practices that lead to misleading or unreliable results.

Open science also allows others to scrutinize the analytic decisions of researchers, such as whether the analysis was planned before or after observing the data, according to a 2018 study.

This allows others to check if they can arrive at the same conclusion as the original research team, and facilitates stronger public trust and support, according to a 2021 UNESCO report.

Here is an example of open science that was able to correct errors in a healthcare study quickly, saving lives! In 2021, a study was published that found that Covid stay-at-home policies did not stop transmission of the virus. The study was highlighted by prominent lockdown skeptics and news sites – swiftly gaining the attention of many people right at a critical time in the pandemic. Here was a scientific research article that said lockdowns don't work! The authors of the study published source code and data with their paper. This allowed others to quickly look at how they arrived at their conclusion. Almost immediately, questions were raised about the paper and within nine months, two papers here and here pointed out major analysis method errors. The original paper was retracted. We all make mistakes. In this case, the paper had major policy implications and because the original authors had practiced open science, the error was rapidly corrected!

3.6.3 Open Science Leads to More Discoveries

The Solar and Heliospheric Observatory (SOHO) has been sending home images of our dynamic sun, opening up a new era of solar observation. It was designed for heliophysics. However, planetary scientists found SOHO useful for its ability to spot comets that pass extremely close to the sun, known as sungrazers. To this day, SOHO is one of the best sources for views of the giant surface explosions regularly produced by the sun called coronal mass ejections, or CMEs, which can hurl a million tons of solar particles off into space. This field of view is large enough to see a sungrazing comet as it sling shots around the sun.

SOHO's great success as a comet finder is, of course, dependent on the people who sift through SOHO's data – a task made open to the world through real-time publicly available data.

A cadre of volunteer amateur astronomers dedicate themselves to searching this data via the NASA-funded Sungrazer Project. While scientists often search the imagery for very specific events, various members of the astronomy community choose to comb through all available imagery in fine detail. Over 2,300 comets have been found, 75% by citizen scientists. This created a great training dataset for algorithms. NASA scientists had algorithms to find comets that they felt that were sufficiently accurate.

In 2022 though, NASA decided to fund a challenge open to the public to develop new algorithms and guess what? New algorithms were discovered along with two new comets!

3.6.4 Quality and Diversity of Scholarly Communications

Furthermore, open science improves the state of scientific literature. Scientific journals have traditionally faced the severe issue of publication bias, where journal articles overwhelmingly feature novel and positive results, according to a 2018 study. This results in a state where scientific results in certain disciplines published scientific results may have a number of exaggerated effects, or even be "false positives" (wrongly claiming that an effect exists), making it difficult to evaluate the trustworthiness of published results, according to a 2011 and 2016 study. Open science practices, such as registered reports, mitigate publication bias and improve the trustworthiness of the scientific literature. Registered reports are journal publication formats that peer-review and accept articles before data collection is undertaken, eliminating the pressure to distort results, according to a 2022 study. Other open science practices, such as pre-registration, also allows a partial look into projects that for various reasons (such as lack of funding, logistical issues or shifts in organizational priorities) have not been completed or disseminated, according to a 2023 study, giving these projects a publicly available output that can help inform about the current state research.

By using openly available tools and making our scientific process and products more openly available, we can ensure that all who wish to involve themselves can take part in the global scientific community.

3.6.5 Key Takeaways: Benefits to Science

- Open science can accelerate scientific discovery. Collective knowledge is not only faster, but more effective than individual efforts.
- Open science allows for errors to be quickly corrected, making science more accurate.
- Open science practices, such as registered reports, mitigate publication bias and improve the trustworthiness of scientific literature.

3.7 Benefits to Society

The mainstream adoption of open science began relatively recently. The potential benefits of open science extend beyond research through contributions to society and policy.

Collaboration, innovation, education, technology advancement, and science-based public policy are all improved by the open availability of research products. Sharing all research products (eg. data, code, results) makes the scientific process more transparent which may help increase public trust in science. Also, open science encourages IDEA (Inclusion, Diversity, Equity, Accessibility), and increases involvement of citizen-scientists and non-experts in the research process. The inclusion of diverse perspectives from an open community invites unique perspectives that contribute to a more robust and often more accurate scientific outcome.

Scientists study issues that affect every aspect of life. Yet, public interest in science remains low due to a lack of trust, understanding, and sociocultural factors. How can scientists expect the public to trust science about complex and often contentious issues, whether it is vaccine development or landing on the moon, if they don't allow the public to see the process and results? Building trust in science is essential to a well-informed society. Open science provides a pathway to do this.

The public who funds government research through taxes should be entitled to its results and data, as long as safety and security are not an issue. Science should be more open to ensure its insights benefit the public who enables it.

Open science introduces more scrutiny into research that helps ensure accuracy and encourages efficiency through open discourse. This approach accelerates the pace of discovery and subsequently the dissemination of results to the public and policymakers.

3.7.1 Open Science Can Accelerate the Pace of Science

Open science practices accelerate the pace of scientific discovery by involving ideas and labor from the broader community. The rapid response to the Covid-19 Pandemic showed Open Science in action to accelerate discovery.

Researchers uploaded the initial genome sequence of SARS-CoV-2 into an open-access database in January 2020, creating a data-sharing precedent and metadata that would later enable insights about new Covid-19 variants. The NIH developed a dedicated platform for sharing research tools for Covid-19 and encouraged investigators to expedite reporting to ClinicalTrials.gov ahead of requirements. Open-science publishing agreements that support evidence dissemination have complemented these practices and policies. One day after the World Health Organization declared Covid-19 a public health emergency, more than 50 academic publishers issued a joint statement committing to open-access policies for Covid-19 research. Support for preprint servers has promoted awareness of research successes and failures, and journals have helped accelerate the distribution of actionable information, including by means of dedicated

Covid-19 web pages, endorsement of preprints, and an emphasis on sharing data with public health authorities.

3.7.2 Open Science is Efficient Science

Open science reciprocates the benefits it provides to researchers onto the communities that scientists hope to serve. Data from one observation or science experiment can have unanticipated uses. In Lesson 1, we discussed an example where the use of radar data for tracking the effect of climate change was used to track bird migration.

Through open science practices, research waste can be avoided, such as unintentional and costly repetition of previous studies, according to a 2020 European Commission report. In the human sciences, this also reduces participant fatigue in the long term. By maximizing what is learned from publicly available data, one does not need to test repeatedly, especially on already vulnerable communities. By "giving away" science, individuals, communities and organizations can more easily adopt research results to inform interventions for their own needs without the knowledge being gatekept by the original researchers and organizations involved. In this way, open science can strengthen the social and economic impacts of scientific results.

3.7.3 Open Science Attracts a Diverse Set of Participant

Image credit: Andy Brunning/Compound Interest. **CC BY-NC-ND 4.0 DEED**

The open sharing of scientific products and processes makes science accessible to everyone. This allows full participation from everyone, and also maximizes the number of people who can benefit from the work.

The best ways to include a diverse group of open science practitioners and stakeholders are to remove existing barriers and design for inclusion. Beyond this, it is important to learn how to communicate effectively with diverse collaborators, people at different skill levels, career levels, backgrounds, and areas of expertise. The ability to build diverse teams is a skill that everyone can learn.

3.7.4 Key Takeaways: Benefits to Society

- The public who funds government research through taxes should be entitled to its results and data, as long as safety and security are not an issue.
- Open science allows for errors to be quickly corrected and accurate results to be built upon by others, impacting policies.

- Open science decreases the unintentional and costly repetition of previous studies, accelerating science that benefits society.

3.8 Lesson 2: Summary

The following are the key takeaways from this lesson:

- Citing the work of other scientists whose work you build upon or reuse supports the community-minded open science practice of using, making, and sharing.
- Doing science openly can boost the visibility of research and lead to more meaningful collaborations.
- Science quality and efficiency is improved when open science best practices are followed.
- Open science helps society by allowing more people to participate in science, which increases the accuracy and impact of results.

3.9 Lesson 2: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/03

What benefits do the individual researchers gain when practicing open science?

Select all that apply.

- Documenting work
- Receiving credit
- Finding collaborators
- Priority access to international scientific conferences
- Securing a higher position in the academic hierarchy irrespective of their publication record

Question

02/03

How does openness improve science?

Select all that apply.

- Accelerated pace of science
- Improved accuracy
- Reduced publication bias

- Fewer discoveries

Question

03/03

What benefits to society does open science bring?

Select all that apply.

- Wider use of results increases return on investments in science.
- Open distribution of accurate results reduces misinformation.
- Open Science facilitates a greater understanding and acceptance of pseudoscientific theories.
- It creates more job opportunities in science-related fields, irrespective of the relevance or quality of the research.
- Promotes the commercialization of academic research, leading to increased product advertisements in scientific literature.

4 Lesson 3: How to do Open Science

4.1 Navigation

4.2 Overview

The ability to discern when and how to share information in an appropriate manner is an essential skill of open science. Practitioners of open science must balance their pursuit to maximize openness while respecting diverse cultures, maintaining security and privacy, and following institutional policies and practices.

This lesson introduces important security and privacy considerations for scientists when sharing information. Next, the lesson discusses how sharing information may impact different communities. Following this, the lesson explains the topic of intellectual property, how it can be protected, and the different types of licenses available to facilitate sharing while ensuring the owner of the information receives credit for their work. Lastly, this covers the effect of rules and regulations set by an organization, grant, or publisher on a scientist's options to make their research open access.

4.3 Learning Objectives

After completing this lesson, you should be able to:

- List reasons information should not be shared due to security or privacy issues.
- Define what intellectual property is and recall the different ways it can be shared openly through licenses or the public domain.
- Recognize sharing policies and procedures of your department, organization, funding agency, and publication in order to make the most responsible science sharing decisions.

The following paragraphs of this section outline key areas of consideration for determining whether or not to make your data openly available.

4.4 Maintaining Security and Protecting Privacy

Previous lessons have showcased a broad range of open science success stories, but we recognize that there are still plenty of valid concerns and unexplored challenges to implementing open science. Open science demands the valuable but complex practices of respecting diverse cultures, maintaining security, and protecting privacy. This lesson presents a strategic approach to making decisions about doing open science in common scenarios. For those scenarios that we cannot foresee, this lesson offers mitigation strategies to help overcome unique challenges with mindful preparation and community support.

Scenario: A Country's Military Secrets or Violates National Interests

When the release of data or research can lead to national security concerns, there are added restrictions around sharing this information. In the U.S., sharing of this type of information often falls under International Traffic in Arms Regulations (ITAR) and Export Administration Regulations (EAR) export control regulations. Sharing ITAR/EAR-regulated data, equipment, resources, or research without clearance to do so can put the country's national security at risk and may bring about both severe criminal and administrative penalties.

Human Patient Privacy

NASA has collected human spaceflight biomedical data since the start of Apollo…

… but the only human data in the Life Sciences Data Archive are from astronauts who signed releases for their data to be public.

In the U.S., health data is protected under the Health Insurance Portability and Accountability Act of 1996 (US-HIPAA) and it is not allowed to be shared without expressed written consent by the patient. As such, health information about astronauts is something NASA protects carefully, working to balance the publicity of the job with regulations and best practices for medical privacy while also enabling peer-reviewed biomedical research.

See this example and more at NASA's Open Science Data Repository.

Respecting Diverse Cultures

Open Science advocates for making research widely available, while also recognizing that there are many reasons why some information should not be released, and that these decisions need to involve the people who provided input and/or could be harmed by the consequences of release.

Indigenous, Cultural, and Conservation Concerns

When considering the impacts of data sharing, it is important to recognize if those affected are equally represented in the discussion. For example, historically excluded communities, the environment, and wildlife are too often not considered when deciding to make research open access.

For example, while genomic research often relies on individual- based consent, it is often used to make decisions that impact indigenous communities without their consent.

Another example of how data can inadvertently impact vulnerable communities is the use of LiDAR by archaeologists to study remote areas. This type of data has the potential to reveal unprotected vulnerable indigenous sites in need of protection.

CARE Principles

The CARE Principles of Indigenous Data Sovereignty are people- and purpose-oriented, and were originally set up to use data in a way that advances data governance and self-determination among Indigenous Peoples. CARE principles can be applied by involving communities or local stakeholders and should be covered at the start of a research project.

Environmental Justice

When sharing your results, are you sharing it with the groups that are most impacted in ways that are accessible to them? When studying the impact or effect on a specific community, it is important to include that community in the design of your work and ensure that the results of the work are accessible - both freely available and understandable – to the communities involved.

Environmental justice is the fair treatment and meaningful involvement of all people regardless of race, color, national origin, or income with respect to the development, implementation, and enforcement of environmental laws, regulations, and policies.

Protecting Endangered Species

Humans aren't the only group that can be negatively impacted by data sharing. Rare and endangered species can also be impacted. For example, the sharing of breeding sites for declining wildlife populations can further exacerbate the population decline. For this reason, rare animals may have their breeding sites kept secret.

4.5 Intellectual Property

4.5.1 What is Intellectual Property?

Intellectual property is the recognition of rights associated with the content created by human intellect. There are several different types of intellectual property and how they are recognized varies by country, type, and timescales.

It's important to understand who has the rights to the content you create. It can depend on a number of different factors. Work that you create may belong to your employer, may be in the public domain, may depend on the license of underlying work, may belong to the publisher of your work, or may be your own intellectual property. Ownership may affect how your work can be shared.

This section provides an introduction to some of the common issues faced by researchers around intellectual property. For instructions specific to your institution, reach out to your intellectual property counsel at your institution for details of how these may affect sharing your scientific work.

4.5.2 Most Common Types of Intellectual Property Protection

Copyright

A copyright protects original works of authorship. This could be artistic or literary works, and also applies to software. In general and if applicable, copyright is automatically applied at the moment of creation with no further registration needed.

Most open licenses depend on copyright. The person(s) who owns the copyright has the right to apply for a license.

Example: An image in a scientific journal or something from the web. Generally speaking, using copyrighted images for teaching and education is considered fair use. However, if that includes posting images to a website, that could be considered a publication and therefore copyright infringement.

Trademark

A trademark can be applied to any content including words, phrases, symbols, designs, or a combination of these things that identifies your product. Trademarks in general are not relevant for scientific purposes.

Patents

A patent is an exclusive right granted for an invention, which is a product or a process that provides, in general, a new way of doing something, or offers a new technical solution to a problem. Patents are another way to make your work open while protecting your intellectual property.

Many organizations have groups that will support the development and commercialization of inventions. NASA's Tech Transfer office is an example of one of these making much of NASA's inventions available for licensing as part of the NASA Patent Portfolio.

Public Domain

In some cases, intellectual property is not protected at all. Public domain is when a creative work has no intellectual property rights associated with it. Some types of intellectual property expires after a certain time scale. Some types of work, such as those created by civil servants in the United States, is not covered by copyright and can appear immediately in the public domain. For others, the creator donates the work to the public domain or intellectual property rights are not applicable.

4.5.3 Why Should You Care About Intellectual Property Policies?

Why should I, as a scientist, care about this? Well, consider what happens to the ownership of your research if you move institutions:

- Can you take your paper drafts, presentations, and copies of publications with you?
- Can you take your data?
- Can you take your software?

Understanding these questions is important to practicing open science and ensuring that your intellectual property is able to be shared widely. Review the image on the right as an example.

Worrying about intellectual property and copyright can seem like an unnecessary detail early on. However, anticipating changes to your situation by ensuring permanent ownership of your work in the planning phase of your research can help you avoid legal and institutional issues later on.

If you submit your manuscript to a publisher that requires that they own the copyright of the work, will you be able to access that paper when you change jobs and no longer have a subscription to that work? Are you able to meet the mandates of your funding agency to openly share your work? Can you reuse the figures that you made in derivative works? Will others be able to access your work? While these may seem like questions you shouldn't have to worry about, it can become very difficult to deal with after the fact.

Example: In scenarios where seeking consent before sharing (or changing sharing conditions), it can be complex to implement the changes. Biopython, an open source biology toolkit, started re- licensing their code in 2016, and are still working on it in 2023, individual contributor by individual contributor.

4.5.4 Licensing

Licensing is a way to help to allow others to reuse your work legally. It is a way to specify under what conditions, if any, others can use, build upon, or distribute your work. It is also a method to ensure that your work is appropriately credited. It is generally illegal and may be a form of academic misconduct to reuse content without a license, even if the content can be

found on the internet. This law protects content creators, just as it protects your work from being used by others without clear permission. Thankfully, it's easy to allow others to re-use your work.

"By applying a license to your work, you make clear what others can do with the things you're sharing, and also establish the conditions under which you're providing them (such as cite you)."

Open Science Knowledge Base

Image credit: XKCD: CC BY-NC 2.5 DEEX

https://xkcd.com/14/

If you don't license your work, others can't/shouldn't re-use it – even if you want them to.

Licenses can be applied to data, code, and reports, or publications, and almost any other "creative" output. There are also several different types of licenses and also the case where no license need to apply:

PERMISSIVE LICENSES

PROTECTIVE LICENSES

PUBLIC DOMAIN

Permissive Licenses allow users a wide range of rights including the ability to use, modify, and distribute the work with no restrictions or very few. Examples of permissive license would be open source software license such as Apache 2.0 or MIT license or the Creative Commons licenses such as Creative Commons Attribution (CC-BY).

PERMISSIVE LICENSES

PROTECTIVE LICENSES

PUBLIC DOMAIN

Protective Licenses are a legal technique of granting certain freedoms over copies of copyrighted works while including some limitations. This may include copyleft licenses, commercial licenses, or other restrictions.

PERMISSIVE LICENSES

PROTECTIVE LICENSES

PUBLIC DOMAIN

Public Domain is not a license, but it is an indication that there are no reuse restrictions on the work. Creative Common Zero is a worldwide public domain mark that indicates that the material is free to use without any restrictions.

More detail about licensing for each of these types of products can be found in later modules including different types of licenses, when to apply a license, and tools for applying licenses. Creative Commons and the Open Source Initiative are two resources with more information on open licenses.

4.5.5 Activity 3.1: To Share or Not to Share

Place the following activities into the Share or Not Share boxes:

Share

Plant genome data from experiments on the International Space Station.

Software that identifies solar flares.

Artificial Intelligence model trained on Mars observations.

Not Share

Export controlled design documentation for a high resolution camera.

Non-anonymized astronaut health data.

Software containing login keys to high performance computers.

4.6 Policies and Practices around Open Science

So far we've discussed situations where sharing may not be appropriate or may need to be handled with care, as well covered intellectual property protections to consider prior to sharing your work. Next, we will explore the importance of policies and practices to practicing open science set by your organization, funding agency, and academic journal. The decision to not release certain scientific results can be a moral and/or legal choice. The next section of this lesson is an introduction on what to watch out for.

4.6.1 Preparing to Use and Make Controlled Research

It is important to plan for the release of your data and results from the very beginning of your research project. Investigate and obtain all permits, approvals, and/or certifications needed to ensure you can share your research products.

Remember: Reputable journals and repositories will reject submissions if compliance can't be documented!

MATERIALS - SHARING AND COMMERCIAL PRODUCT AGREEMENT

HUMAN OR ANIMAL SUBJECT INSTITUTIONAL REVIEW BOARDS

COLLECTING PERMITS

Can be permissive or restrictive.

Many versions available.

MATERIALS - SHARING AND COMMERCIAL PRODUCT AGREEMENT

HUMAN OR ANIMAL SUBJECT INSTITUTIONAL REVIEW BOARDS

COLLECTING PERMITS

Check experiment-specific requirements early.

Be sure to comply with all aspects of ongoing review.

MATERIALS - SHARING AND COMMERCIAL PRODUCT AGREEMENT

HUMAN OR ANIMAL SUBJECT INSTITUTIONAL REVIEW BOARDS

COLLECTING PERMITS

Don't assume collection is allowed just because a sampling location seems unmanaged.

Engage and consult with local communities to ensure their concerns are addressed.

4.6.2 Sharing Controlled Research

As we've previously shown, different kinds of intellectual property are released using different formal structures. For example, text and media products are released under copyright and software is released under a license.

It is important to check with specialist communities when preparing your research plan. Methods for sharing results may follow different standards of practice or may require a special data format for distribution or submission to common repositories.

CREATIVE COMMONS VS. OPEN SOURCE VS. PUBLIC DOMAIN LICENSES

REPOSITORIES

GUID ELEMENTS FOR SELECTING BETWEEN OPTIONS

Can be permissive or restrictive.

Many versions available.

CREATIVE COMMONS VS. OPEN SOURCE VS. PUBLIC DOMAIN LICENSES

REPOSITORIES

GUID ELEMENTS FOR SELECTING BETWEEN OPTIONS

General and discipline-specific options.

Check submission requirements early.

Often have user communities willing to help.

CREATIVE COMMONS VS. OPEN SOURCE VS. PUBLIC DOMAIN LICENSES

REPOSITORIES

GUID ELEMENTS FOR SELECTING BETWEEN OPTIONS

Choose 'supported' versions with active and friendly communities.

Take precautions to reduce security risk.

What are the rules for science? Before sharing, check you have the right to do so:

1. What does your supervisor or Principal Investigator say?
2. What does your grant/contract say?
3. What does your organization say?
4. What does your funding agency say?
5. If you are planning to publish, what does the publisher say?

Remember, sometimes what they say may conflict, for example:

- If your grant / funder says outputs should be open, usually your institute will permit you to share items even if they are normally more restrictive.
- Different types of outputs may have different types of restrictions. (e.g. software or hardware might have one expectation, whilst data might have others).

Universities and other institutions may have OSPOs (Open Source Policy Office) or commercialisation offices. Most institutes will have intellectual property counsel to help answer questions. Librarians are another good resource to consult when looking for advice on sharing. Considering these policies earlier in your research can save you time and energy down the road, which is why…

4.6.3 Early is Better

It is important to think about what policies may affect your research outputs as early as possible so that when you want to share information, you have either already obtained approvals or know where to go to get approvals to share. This ensures that you don't inadvertently share (or fail to share) something that could affect your career, negatively impact others, or pose legal issues.

Remember: You can't unshare something that is already shared! Equally, if your research requires ethical approval or consent to share, this may be harder to gain after you've done your study.

This also helps structure your research, data, and methods in a way that makes it easier for you to share when the time comes.

4.6.4 Reusing Science Ethically - Give Credit!

As we stated previously, licensing helps make sharing your work easier, but it also ensures you retain credit. It's always important to properly source any content you use and remember to only share properly licensed content. Even if a license does not require attribution, providing credit helps increase reproducibility by providing the provenance of your work. This is the norm in scientific communities.

Remember when reusing science:

- Open science is a partnership and giving credit is critical to make it work.
- Consider citing all resources used: datasets, software, infrastructure, etc.
- Hopefully, others will reciprocate when reusing your work. (Scientific ethics dictate they should).

Standing on the shoulders of giants (Public Domain)

Image credit: Library of Congress, Rosenwald. CC-BY

In addition to documenting your own research, giving credit to all contributors strengthens the practice and community of Open Science. As researchers gain confidence in the benefits of cooperative research, they will in turn give credit to contributors that might otherwise have gone unacknowledged.

4.6.5 Other Reasons Not to Share

This lesson has only covered common scenarios of when and how to share science. Regardless of the situation, it's important to consider the implications of sharing information and those who could be negatively impacted before deciding if or how the information should be shared.

In order to practice responsible open science, careful attention should be given to how data is anonymized and how sensitive information is removed from it in order to safeguard people's identity and prevent breaches of privacy. The misuse of private data and illicit means of collection is an issue for every sector, not only science.

4.6.6 Activity 3.2: Not all Science Can, or Should, be Open All the Time

In this activity reflect on your answers to the following questions:

- What are some reasons you would NOT want your research to be open?
- How would you balance openness with privacy/security/control?

4.7 Lesson 3: Summary

In this lesson, you learned:

- Situations when it may be inappropriate or harmful to share your data or research. These include maintaining security, protecting privacy, and respecting diverse communities.
- What intellectual property is, who owns it, and how it is protected through licenses.
- Various organizations within science (e.g. universities, publications, funding agencies, etc.) may have their own individual sharing policies that are best to consider at the beginning of a research project to avoid any potential pitfalls along the way.

4.8 Lesson 3: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/03

What type of Intellectual Property protection would you need for an image you generated from your own data to display your results?

- Patent
- Trademark

- Copyright

Question

02/03

Which of the following is NOT a common type of license:

- Copyright
- Permissive
- Creative Commons

Question

03/03

Which of the following should be considered when sharing information?

- What does your funding agency say?
- What does your organization say?
- How are affected communities impacted by the release of the information?
- All of the above.

5 Lesson 4: When Not to be Open

5.1 Navigation

5.2 Overview

In this lesson, you will consider potential barriers to adopting open science practices. Barriers can come in the form of personal fears, as a result of misaligned social challenges, or institutional/infrastructure barriers. We begin with an exercise to identify your own concerns or fears around adopting open science. This leads into a discussion about common barriers and mitigation strategies.

5.3 Learning Objectives

After completing this lesson, you should be able to:

- Recognize your own fears and concerns for adopting open science, and list mitigation strategies for overcoming them.
- List common barriers to practicing open science that occur from misaligned incentives and mitigation strategies.
- List several social challenges that can arise when practicing open science and strategies for communicating effectively to overcome differences in perspective.
- List several institutional and infrastructure barriers to doing open science and mitigation strategies where available.

5.4 Common Fears Around Openness

5.4.1 Activity 4.1: Self Reflection on Open Science Concerns

Take a moment to think about what fears or concerns you have around adopting open science. These could be concerns you have experienced in your work, or fears you have for being more open moving forward. There are no wrong answers here – this is a time for you to reflect on what might be keeping you back from doing open science.

Some Fears Around Adopting Open Science Practices

Now that you've reflected on some of your concerns or fears around open science, below we have listed a few common fears of doing open science and some potential mitigation strategies. Even if you personally don't have this fear, it can be useful to think about the different concerns that others may have to better understand and even help others address them.

Table: Common Fears About Open Science

Fear

Discussion/Mitigation

Mistakes: What if my work is wrong?

It can be intimidating to share your research materials publicly, because someone might find a mistake. But isn't it better for science if we can quickly find and fix mistakes? Peer review is a core pillar of the scientific method, and is a mechanism for others to help find and correct mistakes. To make this work, we will need to be more open to finding and fixing mistakes. It's true that in many science communities, a mistake is considered a failure. However, open science policies aim to change the perception of mistakes from that of failure to a step in the discovery process that can be aided with open community feedback.

Scooping: What if someone re-uses my work and gets the credit?

Yes, this can happen.

Depositing your work early and making it citable are ways to establish your work.

This serves as evidence for when you started working on it and makes it easier for others to cite you. Details of how to do this are provided in the following modules.

In many fields, if it is clear that someone is actively working on a problem, the decision to scoop that work may have a short term gain but long-term loss. In science, reputations are very important and being collaborative generally leads to increased career successes. Read more about scooping here.

Misinterpretation of my work.

This can happen regardless of the form or openness of your work - many publications have ended up being misinterpreted.

Openness does help to provide further context of the work. Documentation of your research plan and software management practices allow others to understand your work fully, and thus help reduce the risk that others will misinterpret your work. For example, if you share code, you can include a description of what the code does, along with brief usage instructions and examples. In Module 4, we will discuss proper data and code documentation that can help reduce misinterpretation.

My work will be used, but not cited.

Science ethics dictates that you should be cited if your work is used. Part of open science is valuing all steps of the scientific workflow, and encouraging researchers to cite code, data, or other non-published articles. Make it easy for others to cite you by adding a digital object identifier (DOI - discussed later in the course) to your research product. Remember to cite others' materials, so you're not adding to the problem.

Data is too sensitive to share.

Following appropriate anonymization or using controlled access can address this concern.

I don't want to maintain or update my work.

Sharing what you did allows others to reproduce, replicate, and build upon your work. That doesn't mean you have to maintain it for the rest of your life, or even at all. If you don't plan to maintain your code, it is still recommended that you share the code publicly and archive it. By adding appropriate licensing, documentation, and contributing guidelines, you can make it clear how long you plan to keep your materials maintained (if at all). In fact - others might help maintain it for you!

My work won't be useful to anyone else.

You never know how materials might be used. Individuals who contributed to all different types of software projects ended up helping NASA land a rover on Mars!

Partially drawn from Malvika Sharan's "Ten Lessons Against Open Science You Can Win"

Some of the fears listed above are not unique to open science and can occur in closed scientific systems. For example, scooping and reuse without citation are both examples of scientific misconduct that can happen in closed science scenarios. Open science practices can provide more avenues for recourse, such as making a preprint available or giving your data or code a DOI and license. Having more of your work shared in citable ways gives you more power to prove when misconduct has occurred.

Another example of a fear that occurs in both open and closed spaces is the commitment to maintaining your work beyond publication. Maintenance is a consideration regardless of whether your work was shared - you need to decide how long to store your data and code for yourself in order to reproduce your work, should any questions arise even after publication (we

cover sharing and archiving data and code in later modules, Open Data and Open Code.) By sharing your research materials, you may actually increase the longevity and impact of what you've done if others find your materials useful and help maintain and build on top of them.

We recognize that this is not an exhaustive list of concerns and fears toward adopting open science. We have developed this module of the TOPS curriculum to provide guidance and instill confidence for researchers who intend to do their work more openly moving forward.

5.5 Misaligned Incentives

Previously, we discussed some fears and concerns of adopting open science. In this section, we discuss barriers that block participation in open science that stem from misaligned incentive structures. These all relate to scientific incentives for individuals and organizations, and are not aligned with open values.

We distinguish between concerns and fears; those associated with changing the culture of how we do science; from the structural barriers that block researchers' abilities to adopt open science practices. We recognize that there is overlap in these categories, but this framing might be useful for understanding what we have control of as individuals, and where we need to encourage more structural changes to our scientific ecosystem.

5.5.1 Overview: Misalignment of Incentives

Incentives can come in many forms, but most in science involve proposal funding and career advancement. In both of these cases, metrics are used for measuring scientific success (e.g., publication and citation count, as discussed earlier in this course). These current metrics do not capture the entire impact of activities that scientists spend their time doing. Below, we present a few examples of misaligned incentives. While there aren't perfect answers for overcoming these yet, initiatives like DORA and COARA are actively working to update these metrics that define what success means in science, and it will take community action to ensure that open and inclusive practices get the merit they deserve.

5.5.1.1 Challenge: Overvaluing Novelty

Awards (for example prizes or funding) are often given to those who make a big, new scientific discoveries or who create a new, exciting tool. This practice overlooks the community that wrote code, curated datasets, maintained fundamental existing tools, and many other important steps that enabled these novelties.

Prizes often disincentivize crediting a team, since only one or a small group can be awarded a prize (for example, a Nobel Prize can be awarded to up to 3 people only). This emphasis on

novelty and the individual are starting to change, with awards being offered to groups (e.g., The White House Office of Science & Technology Policy Open Science Recognition Challenge) and addition of funding solicitations offered for maintaining tools and infrastructure. However, it will take time for these changes to become the norm.

5.5.1.2 Challenge: It Takes More Time to be Open

Doing open science often requires more time and effort from researchers to start and maintain. For instance, it can take significantly more time to document and clean code to a degree that the public can easily understand and use it. At the moment, the scientific system doesn't always reward extra effort like this, which can make it difficult for individuals to spend their time on open activities because it takes time away from starting their next paper. After all, published papers are the main currency of the current scientific system.

Updated metrics of success can help to incentivize individuals to do their work openly. The science community is currently in a transition phase where new metrics are being developed, but the old metrics still dominate in many fields and organizations. It's important for researchers to recognize that they might not be able to achieve complete openness until the system and culture shifts.

5.5.2 Activity 4.2: To be Open or Not to Be...

In this activity, reflect on your answers to the questions and then compare your thoughts to the key takeaways.

Image credit: NASA 2023 @ Stennis Space Center.

Conferences are open places – most of the time. Think about who can attend a conference. How open/closed is it?

Publications can have both open and closed elements. How is it open?

5.6 Social Barriers

5.6.1 Challenge: Collaboration & Community - Open community members don't always agree with one other

Meaningful collaborations across diverse communities can require additional time and effort to coordinate across groups and to address conflicts. While interacting with the community can be

one of the most fulfilling things about Open Science, it might also be a source of disagreements about the direction of the project or how it should be used. That's where licenses and codes of conduct come into play. Clear rules for community- and colleague- interactions and use of resources provide a framework to make decisions in a fair and agreed-upon manner. This can all take additional time, especially at the beginning of a research project, but can save time and headaches down the road.

5.6.2 Strategies for Communicating Across Differences

These are ways you can encourage openness in your discussions around research. For in-person sessions, it's good to encourage discussion of these strategies:

- Presume that everyone you work with is doing the best they can at the time.
- Attempt collaboration before conflict.
- Listen carefully and actively.
- Encourage other people to listen as much as they speak.
- Practice empathy and humility.
- Ask questions that seek to understand your colleagues' context.
- Participate in an authentic and active way that supports the health and longevity of your community.
- Exercise consideration and respect in your speech and actions.
- Treat other people's identities and cultures with respect: e.g., make an effort to say people's names correctly, and refer to them by their chosen pronouns.
- Be mindful of your surroundings and of your fellow participants, and take action if you notice a dangerous situation or someone in distress.

5.7 Institutional and Infrastructure Barriers

5.7.1 Institutional Barriers: Institutions Often Move Slowly

Institutional barriers to the researcher or practitioner present an additional challenge to adopting open science practices. Researchers interested in adopting open science practices might lack support from their department or project supervisors. The budget, resources, or time in a project cycle might be insufficient to practice open science. Institutions might not recognize open science practices in recruiting, training, or promoting in the organization. Even if organizations show interest in moving toward open science, they can move slowly when setting up new systems of support.

In these situations, there isn't always an obvious mitigation strategy. While we encourage individuals to practice open science, there may be aspects that just aren't feasible at this point in time without spending a lot of extra time and effort, time that may not be recognized

or supported by your institution. It's best to work within the bounds of the system you are in, and while the entire scientific community is in a transition phase to being more open, it may be that it doesn't make sense to be open in every way until the institutional barriers are lowered. That said, the more individuals that push for openness, the - more it will become part of the scientific mindset, and the more likely our organizations are to recognize and support our efforts.

5.7.2 Tools & Infrastructure

5.7.2.1 Do the right tools and infrastructure exist to support my work?

There are many tools and resources for making our code, data, and results more open, but the required infrastructure is still being built, and may not be in place yet to support open science in each discipline. This is where community input can be helpful. Perhaps there is a community already working on implementing the infrastructure you need. If not, you can start discussions at conferences or on open online forums to help organize the creation of the tools and infrastructure you and your community need to effectively do open science.

5.7.2.2 How can I get around institute-specific infrastructure when trying to collaborate with people outside my organization?

Some of our infrastructure (like our computing platforms) are institute-specific, which can be a barrier to collaboration outside of our organization. However, by planning for open collaboration from the start, you can minimize these barriers. For example, you can use freely available tools like GitHub and Google Docs for communication and coordination, even if the computing facilities are institute specific.

5.7.3 Open Science is Worth the Effort!

While there are many challenges to the adoption of open science, we believe that its benefits and its ethical imperative to the self and to scientific communities, citizens, and policy-makers outweighs the cost of barriers. In addition, the recognition of barriers and areas for caution provides a first step towards resolving them.

5.8 Lesson 4: Summary

The following are the key takeaways from this lesson:

- There are valid concerns and fears around making our science more open, but there are often specific open science practices that can help to mitigate these fears.

- The misalignment of incentives creates real-world challenges that act as barriers to adopting open science practices. There are ways that individuals can minimize or work with these barriers, as well as organizations and groups that are actively working to update the incentive structure.
- Working openly and collaboratively has its challenges, but there are some strategies for communicating across differences.
- There are also institutional and infrastructure barriers to adopting open practices, but by using general tools and infrastructure we can minimize some of these challenges.

5.9 Lesson 4: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/05

Match the open science concern with a potential mitigation strategy.

Share your work early and make it easily citable

Someone might reuse my work and get the credit

Add clear licensing, documentation, and contributing guidelines

Someone might misinterpret my work

Provide appropriate documentation for your research products

I don't want to maintain my work

Question

02/05

Read the statement below and decide whether it's true or false.

Open science can take more time and resources.

- True
- False

Question

03/05

Which of the following are strategies to encourage openness in your discussions?

Select all that apply.

- Attempt collaboration before conflict.
- Speak loudly.
- Listen carefully and actively.
- Presume that everyone you work with is doing the best they can.
- Use jargon specific to your discipline.

Question

04/05

Read the statement below and decide whether it's true or false.

All the needed infrastructure is in place to support open science across all disciplines.

- True
- False

Question

05/05

There are times when open science can be a source of disagreements about the direction of the project or how it should be used. What factors should project creators consider to overcome those obstacles?

Select all that apply.

- Governance
- Planning
- Licenses
- Codes of Conduct
- Setting expectations for community interactions

6 Lesson 5: Planning for Open Science: From Theory to Practice

6.1 Navigation

6.2 Overview

This module is nearly over, but there's so much more information available about open science – so our last lesson is for everyone who wants to learn more. In this lesson, you review ways to start your journey with open science including a list of resources that you can use now.

6.3 Learning Objectives

After completing this lesson, you should be able to:

- List considerations to include in a planning for open science and define an open science and data management plan (OSDMP).
- Describe the different parts of the scientific workflow and how open science can be integrated into it.
- Differentiate real world examples of how a team can use open science.
- List four steps that anyone can take to be more open.

6.4 Planning for Open Science

It is important to think about, discuss, and plan for desired outcomes and processes when you begin your research. Learn about where the best repositories are for your materials; discuss credit and authorship for each separate open science output, and start using open science tools to organize your work. Reach out to repositories in your discipline and institution (usually library) for help. Including this information in your plans will make you more likely to receive funding.

Planning for outputs in advance includes:

- Speaking about it and organizing with your research team
- Deciding which tools to use
- Thinking about authorship and credit
- Engaging with relevant stakeholders and research partners, for example, industry, around open science
- Identifying repositories for software and data
- Identifying journals (or other outlets) for publications
- Highlighting these approaches in your grant and much more

In reality, there is an exploratory stage where sharing one's product may not be part of the plan. During active research and data exploration, data, code, and ideas may be created and deleted even daily. It may not be efficient to spend time making these fully open (eg. creating DOIs, documentation) because you are just exploring! Still, one may choose to make their code public through this process (it should be in some version control repository anyway, no harm in making it public). Part of this planning is beginning to think about what would be valuable to science and figuring out how you might share it.

It is important to discuss open science with your research team, lab, group or partners regularly. Much of responsible open science may seem to be related to outputs – such as data, software, and publications – but preparing and organizing work for these in advance is critical. It is much more difficult to follow leading practices for these at the end of research, in the 'afterthought' mode. Open science is both a mindset and culture that starts when you begin a project.

6.4.1 Open Science and Data Management Plans

Federal agencies and funders consider data management crucial for open science because it ensures that research data is well- organized, accessible, and preserved. In recent years, many have included a requirement as part of proposals or projects plans for an Open Science and Data Management Plan (OSDMP). The OSDMP includes a description of the resources to be used, the products that will be created, how they will be shared, and who will be responsible. These plans can include the data, software, publications, and project governance.

Open science and data management plans are essential because they enhance the credibility and reproducibility of research by ensuring that data is well-documented, organized, and preserved over time. Effective OSDMPs can have the following benefits:

Transparency

Not only builds trust in scientific findings but also allows other researchers to validate and build upon them, fostering a culture of openness and cooperation.

Effective

Data management can lead to more efficient and cost-effective research processes. By reducing the time spent searching for and organizing data, researchers can dedicate more time to analysis and interpretation, potentially accelerating the pace of discovery and innovation.

Reproducibility

A key tenet of the scientific method is reproducibility and a well developed OSDMP helps ensure that others are able to validate your results.

Preservation

The research produced by federal funding represents a significant investment and it is important that research is saved for future generations to access and understand.

Inclusive

OSDMPs can include research tools and processes that can significantly improve research outcomes through collaboration and consultation.

Learn more about OSDMPs in Module 2.

6.4.2 An Open Strategy

In today's world, many foundations and agencies that award research grants increasingly expect proposals to include an open science strategy. By including an open science strategy document in your scientific plan, you ensure accessibility and openness in each step of your workflow. Conclude your comprehensive plan with clearly defined steps to make research outputs easily accessible and openly available. The steps identified in your strategy should be integrated into your everyday scientific processes and practices.

6.4.2.1 Requirements

Every major research foundation and federal government agency now requires scientists to file a data management plan (DMP) along with their proposed scientific research plan. Some ask for additional details on software/code and publications.

6.4.2.2 Include Entire Data Workflow Details in the Plan

Describe your management workflow for data and related research. Other elements, such as code or a publication, have their own lifecycle and workflow which needs to be in the plan.

6.4.2.3 Include Open Terminology and Concepts

Plans that are successful typically include clear terminology about how information is made findable, accessible, interoperable, and reusable. This can include licenses, repositories, formats, and governance of the project.

6.4.2.4 Preservation

Research materials are valuable and reusable long after the project's financial support ends. Reuse can extend beyond our own lifetimes. Therefore, researchers must arrange steps for preservation and accessibility to ensure work is not lost after a research interaction ends.

6.5 Designing for Openness

6.5.1 Open Science Applies to the Entire Workflow

Open Science Workflow Phases Source: Opensciency

Regardless of your science discipline or the methodology that you use, the workflow remains relatively the same. It has a planning phase, an implementation phase and a release phase. Within these phases, there are milestones that vary depending on the workflow you follow. For the purpose of our discussion in this lesson, and the other modules in the curriculum, we have adopted the scientific workflow with general milestones described in the Opensciency curriculum. The details in your workflow may vary, but the overall concepts are the same. What is relevant here is that when adopting open science, it permeates all phases of the workflow. You prepare for it in the planning phase but then continue to integrate the principles of it throughout the implementation and release phases.

Products created throughout the scientific process are needed to enable others to reproduce the findings. Researchers who wish to make their results reproducible must make key elements of their study openly available for others to test.

Open Science Workflow Products Source: Opensciency

Continuing through the workflow, this updated diagram now shows the types of scientific products that are created at each milestone. The specialized products that you create may vary or be completely different, but the focus on discovery for the public remains the same. Any type of products you create can be modified to support the principles and concepts of open science. Where and how to integrate open science concepts into your products is the purpose of this Open Science 101 curriculum.

6.5.2 Use, Make, Share

The idea that open science can impact your entire scientific workflow may seem overwhelming and unachievable, but remember, open science occurs across a spectrum – even small steps towards openness lead to more accessible, inclusive, and reproducible science. And the Open Science 101 curriculum is here to help lead through this process.

In this section, we introduce the "Use, Make, Share" framework that can start to gradually increase your adoption of open science depending on the nature and scope of your project. Throughout the course, we will explore how this framework can be used to make your science more open!

6.5.3 What Resources Will You Use?

There are already many open science resources for you to use! Open science already has a long history. For example, the act that created NASA mandated sharing of its discoveries with all of humanity and NASA has been sharing its data openly on the internet since the 1980's. Now, there are already over 100 Petabytes of openly available NASA data for you to search, download, and use and examples of these services are provided in Module 3. Technology and practices have been developed around code that make it easy to collaborate on building complex solutions, and examples are given in Module 4. A range of services make it easy to share and discover open access publications and these are discussed in Module 5.

In Module 2, we will introduce you to some of the tools that not only make open science possible, but also easy.

6.5.4 What Outputs Will You Make?

Throughout the research process, there will be different products and results produced. These can range from data sets, samples, code, reports, manuscripts, conference proceedings, blog posts, and videos. Each of these have different considerations about how to make them including how they can be made in open and collaborative ways.

There are also different ways to run a scientific project. Is your project going to be open from inception or open at publication? There are valid reasons for both approaches, but generally the earlier you are open with data, code, and results, the more opportunities there are to grow collaboration networks and build with others (which is quite fun). Often researchers choose to be open within their project teams during development, exchanging data, code, and results, but then only sharing with the world once they feel they have a result they can trust. While this approach has been the cultural 'norm' within many communities, this is changing as groups grow more comfortable with openness earlier in projects and experience valuable contributions from others and build new collaboration networks.

Modules 3, 4, and 5 will discuss how to make your data, code, and results open.

6.5.5 How Will You Share?

Image Credit: Freepik.com

Where you choose to share your research materials and results will have a large influence on its impact – how easy it is for others to find it, how long it is available, and how easy it is to reuse.

Will you share data in a file filled with columns of unlabelled numbers without any units or explanations or will it be in an open, standard format and following the Findable, Accessible, Interoperable, Reusable (FAIR) principles? Module 3 has more details to help you better understand how to share your data and explains ideas like FAIR and best practices in sharing data. This includes different considerations for where to share your data as well so that it is both accessible and preserved.

For software, since it is often updated and changed, many researchers first share it on a version control platform like GitHub or GitLab but then archive a version of it in a repository that has long-term preservation capabilities – more on this in Module 4!

For results, open access publications and preprint servers are common locations to share. Module 5 discusses all these options.

6.5.6 Activity 5.1: Use, Make, Share

Take a moment, to answer the following questions on your current research or on research that you would like to do:

- What data, software, or publications do you currently use or would like to use? Are they open or closed?
- What are the tools and processes that you currently use? Is it easy to include others in collaboration?
- How is your work shared or planned to be shared? Can anyone access your results?

6.6 Case Study: The Outcomes of an Open Plan

One of the first discoveries from the James Webb Space Telescope was the first detection of Carbon Dioxide in the atmosphere of a planet orbiting another star. This discovery was enabled through the open science principles adopted both by the project and the team.

Image Credit: @AdobeStock 2023, dimazel

Figure Credit: NASA, ESA, CSA, Joseph Olmsted (STScI)

This was conducted as part of the JWST's Early Release Science (ERS) Program as part of the JWST ERS Exoplanet Transiting Community program (ERS-TRANSIT). This is some of the earliest science data taken with the facility that was made openly available. The team, though, began their work years before the observations and included open science into every step of the process. The team worked in an open format from ideation, to analysis, through to publication and communication.

Let's walk through and see what open results were in fact produced. While doing so, let's take a look at what the advantages of doing so at each stage were.

6.6.1 Planning for Open Science

Opportunity

Create a Governance Plan: An open code of conduct and publication policy highlights the rules of engagement of the final result.

Benefit

Onboarding of new members and facilitated collaboration.

Result

Code of Conduct and Publication Policy.

Availability

On team webpage and GitHub.

One of the most important parts of starting a project is thinking about who is going to be working on it and how they will work together. Before samples are collected, before data is downloaded, before code is written – how will you all work together, what are the roles and

responsibilities, and how and when will you share any materials. That was a key part that this JWST project got right.

The initial team, during the planning phase, developed and openly published information in the form of the code of conduct and the publication policy.

Basically, you are welcome to work with us, but here are our rules. And then the lead scientists regularly talked about this with the team, especially as it expanded so everyone knew what was expected of them and what they got in return (credit!).

The outcome and benefit of publishing this information was the addition of new members to the team, and an agreed to and established collaborative and inclusive culture among the team.

The team grew to almost 400 people, all working together, all knowing what to expect and this created trust.

Read more about collaboration documents and credit on OpenSciency.

CLICK TO LEARN

6.6.2 Open-Source Software**

Opportunity

Collaboration on an open software data-processing pipeline

Benefit

Decreased duplicative efforts, contributors get credit for their work, and accelerated the data wrangling process

Result

Data processing pipeline

Availability

Code on GitHub, released on Zenodo, documents released in Journal of Open Source Software (an open access journal)

Like most data, JWST is complicated and it needs processing and data pipelines. The precision needed to achieve this type of detection requires detailed analysis of the observations and a wide range of expertise.

During the implementation phase, the team collaborated on creating the data processing software together, so that everyone would benefit. Imagine the wasted effort if all 400 people had written the software themselves. The benefit and outcome was that by collaborating on this effort, the team decreased duplicative efforts, contributors got credit for their work, the

software was more accurate, and this effort accelerated the data wrangling process. EUREKA!, the software created, was created openly with documentation and published with peer-review of the software package.

But they didn't have to to start from scratch! The ERS-TRANSIT team was able to build on the work of others. The software built on the JWST pipeline software developed openly by the JWST mission team. Furthermore, they were able to build on a much larger open source software ecosystem using python and Astropy.

6.6.3 Open Access to Results

Opportunity

Research products recorded in public archives and made openly available

Benefit

Allows collaborators to receive individual credit for their contributions, which provided greater incentive for participation

Result

Intermediate products, models, and final data

Availability

All aggregated into Zenodo Community with individual contributing authors and DOIs

During the release phase, the team labeled research products adequately for reuse and reproducibility and published in public forums/repositories. The outcome and benefits included collaborators receiving credit for data, software, and other digital research products that benefit the scientific community. Data and software were archived openly on the general data repository Zenodo and the publication was made available as a preprint and an open access publication.

Making your results open also opens you up to clearer ways of receiving credit and can also reduce the risk of scooping (each result can be individually referenced as soon as it is made available). Applying reproducibility practices separately on different parts of the project such as preparation documents, datasets, software, and reporting allows other researchers to test and reuse your work in their research, and your research will be cited more often, thus bringing fair recognition for your work. Collaborators can get more motivated to contribute because they can easily get recognition in terms of authorship for their contributions made for each one of the research outputs generated.

Read more about making results open throughout the research process on Open-Sciency.

CLICK TO LEARN

6.7 Steps to Continue Your Open Science Journey

Here we will explore the next steps to open science that everybody can take. The thought that open science can impact your entire scientific workflow may seem overwhelming and unachievable, but this is not the case. You can start slowly and gradually increase your adoption depending on the nature and scope of your project. Here are a few immediate ways that you can start engaging in open science.

6.7.1 Where to Go From Here

- Get involved: Become part of an open science community in your sector.
- Start using/sharing the open science tools of your community.
- Learn how to use/archive data in repositories and community tools and resources.
- Concise statement of the Ethos of Open Science: Find, collaborate, and share!

6.7.2 Identify Your Open Science Communities

Here are the steps you can take to find your own science community:

- Talk with your colleagues.
- Read your field's literature.
- Run searches, in general and discipline-specific areas.
- Investigate online communities encouraging open science, such as:
  - The EU's 'Foster Open Science' program
  - The Turing Way online manual
  - FORRT

Join open science communities. There are generic ones as listed here or you can seek out communities that are not only within your domain but also within your geographical area.

- TOPS GitHub discussion board
- Opensciency online open science community list

6.7.3 Explore Open Repositories

There are many repositories that host open data, software, and results. We share many of these resources in the later modules, but here are two NASA repositories that allow you to search for existing data collections that might be relevant to your interests.

- Science Discovery Engine
- https://data.nasa.gov/

6.7.4 Four Steps to Open Science that Anyone Can Take

1. Keep seeking best practices for open science, and develop plans to be more open in your science or research.
2. Think about all the different types of reviews you are involved with, and how to improve them with a goal of openness.
3. Ask colleagues about open science activities, and award credit for them in evaluations.
4. Engage with underrepresented communities to ensure science encourages a more equitable, impactful, and positive future.

6.7.5 Continue Taking TOPS Open Science 101

The TOPS Open Science 101 curriculum is a good place to go for a more in-depth introduction to the various elements of Open Science – each of the next 4 modules provides details and practical exercises to help participants develop a better understanding of that specific topic.

But, if you want to also do an in-person or virtual workshop, we got you!

Self-Paced Online Course

Online Workshops

In-person Workshops

Images Credit: freepik.com

TOPS is coordinating online and in-person open science workshops!

These events provide an opportunity for you to take Open Science 101 with others and engage with members of the open science community! To learn more go here!

6.7.6 Additional Resources

In addition to the resources listed elsewhere in this training, the below community resources are excellent sources of information about Open Software.

References and Guides

- OpenSciency
- Turing Way handbook to reproducible, ethical and collaborative data science

6.8 Lesson 5: Summary

There is no one way of doing open science, and any steps you take to make your science more open are extremely valuable, especially as we transition to a more open scientific ecosystem in the future. We want people to be able to identify the most important things they "can" openly share, but with the ultimate goal of complete openness.

6.8.1 What Have We Covered in this Module?

- Preparing and organizing in advance are crucial components for ensuring the effectiveness of open science work.
- Open Science and Data Management Plans (OSDMP) provide a plan for how open science is integrated into a project including the sharing of data, software, and results.
- Designing for openness is a critical aspect of making sure that open science is integrated into the entire scientific workflow from start to finish. This includes resources that can be used, products that will be made, and how the science will be shared.
- Open Science is already happening - there are already teams conducting their research openly and many resources that can be used to make your research more open.
- There are more opportunities to participate and learn about Open Science – this is just the start!

6.9 Lesson 5: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/04

Applying reproducibility practices separately on different parts of the project allows other researchers to do what?

Select all that apply.

- Expand research focus
- Consider opportunities for more collaborative advantage
- Test and reuse your work in their research
- Cite your research more often to bring fair recognition to your work
- Understand new discoveries

Question

02/04

Read the statement and decide whether it's true or false.

The Open Science and Data Management Plan (OSDMP) describes how the scientific information that will be produced from scientific activities will be managed and made openly available.

- True
- False

Question

03/04

Read the statement and decide whether it's true or false.

Before starting a project, it is important for researchers to consider how they will collaborate with other researchers, what are the roles and responsibilities, and how and when they will share any materials.

- True
- False

Question

04/04

Which item is NOT one of the four steps to open science that anyone can take?

Select all that apply.

- Engage with underrepresented communities to ensure science encourages a more equitable, impactful, and positive future.
- Ask colleagues about open science activities, and award credit for them in evaluations.
- Think about all the different types of reviews you are involved with, and how to improve them with a goal of openness.
- Create a new journal that requires an expensive subscription to read.
- Keep seeking best practices for open science, and develop plans to be more open in your science or research.

6.10 The Ethos of the Open Science Summary

Congratulations! Now you should be able to:

- Explain what open science is, why it's a good thing to do, and list some of the benefits and challenges of open science adoption.
- Describe the practice of open science, including considerations when writing a management plan and the tasks in the "Use, Make, Share" framework.
- Evaluate available options when determining whether research products should or should not be open.
- List ways to connect with others who are part of the open science community.

6.10.1 Further Resources

Masuzzo and Martens (2017). Do you speak Open Science? Resources and tips to learn the language

CLICK TO LEARN

Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data. Publishing version B1.0, Force 11

CLICK TO LEARN

Fecher and Friesike (2014). Open Science: One Term, Five Schools of Thought

CLICK TO LEARN

EC Working Group on Education and Skills under Open Science (2017). Providing researchers with the skills and competencies they need to practise Open Science

CLICK TO LEARN

"What is Open Science"

CLICK TO LEARN

Resources for Earth Science

CLICK TO LEARN

Part II

OS101 Module 2: Open Tools and Resources

Welcome to Open Science 101: Open Tools and Resources

About This Module

This module is designed to help you get started on your journey to practicing open science. It offers an introductory view of the concepts and resources that are fundamental to open science. The bridge between the concepts and the practice of the concepts is something called the use, make, share framework. There are many methods and models that define how to get started with open science. The use, make, share framework was constructed to help you immediately assign purpose to the concepts and tools that are covered in this module as well as in the entire Open Science 101 curriculum. All of the information that you learn here will be addressed in more detail as you participate in other modules but can also be applied immediately after completing this module.

Module Learning Objectives

- Define the foundational elements of open science, which includes research products, the "use, make, share" framework, and the role of an Open Science and Data Management Plan.
- List and explain the purpose of resources used to discover and assess research products for reuse, including repositories, search portals, publications, documentation such as README files, metadata, and licensing.
- Develop a high-level strategy for making and sharing data that employs the FAIR principles, incorporates a data management plan, tracks data and authors with persistent identifiers and citations, and utilizes the appropriate data formats and tools for making data and sharing results.
- Describe the software lifecycle and design a high-level strategy for making and sharing software that considers the the use of a software management plan, the tools needed for development including source code, kernels, programming languages, third-party software and version control, and the tools and documentation used for publishing and curating open software.
- List the resources for sharing research products including preprints, open access publications, reference management systems, and resources to support reproducibility.

Key Terms

These key terms are important topics for this module. Select the term to see the description.

Virtual Machine – A computing environment that replicates the functionality of a physical machine but at a higher level of abstraction on a computer. This allows the specified virtual machine's resources to be more flexible and compartmentalized.

Metadata – Information about the data that provides additional details and context.

Data Repository – An enterprise data storage entity (or sometimes entities) into which data has been specifically partitioned for an analytical or reporting purpose.

Computing Environment – A platform that provides necessary software dependencies, a development area, and connections to computational resources to facilitate running code.

ORCiD – A numeric code used to uniquely identify authors and contributors of scholarly communication. Researchers provide an ORCiD for publications and association memberships. ORCiD is also an international, interdisciplinary, open, non-proprietary, and not-for-profit organization created by the research community for the benefit of all stakeholders including ours and the organizations that support the research ecosystem.

Persistent Identifiers (PIDs) - A long-lasting digital reference to an entity.

Digital Object Identifiers (DOIs) – A string of characters standardized by the International Organization for Standardization, assigned to a piece of digital content, that points to the digital location of the content.

Navigation

Lesson 1: Introduction to the Process of Open Science

Lesson 2: General Tools for Open Science

Lesson 3: Tools for Open Data

- Introduction to Open Data
- FAIR Principles
- Tools to Help with Planning For Open Data Creation
- Tools to Help with Using and Making Open Data
- Lesson 3: Summary
- Lesson 3: Knowledge Check

Lesson 4: Tools for Open Code

- Introduction to Open Code
- Tools for Version Control
- Tools for Editing Code
- Additional Tools
- Lesson 4: Summary
- Lesson 4: Knowledge Check

Lesson 5: Tools for Open Results

- Tools for Open Publications
- Tools for Reproducibility
- Additional Tools for Open Results
- Lesson 5: Summary
- Lesson 5: Knowledge Check
- Open Tools and Resources Summary

7 Lesson 1: Introduction to the Process of Open Science

7.1 Navigation

7.2 Overview

In this lesson you review the definition of open science and several other common terms including research products, data, software, and results. In addition, you will read examples that demonstrate how these open science tools are used in practice. The lesson wraps up with an example of how one group openly shared their data, results, software, and paper.

7.3 Learning Objectives

After completing this lesson, you should be able to:

- Define common types of research products including data, software, and results.
- List common ways to share data, code, and results while practicing open science.

7.4 Definition of Open Science and Research Products

7.4.1 What is Open Science?

"Open Science is the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility, and equity."

The White House Office of Science and Technology Policy (OSTP) and the National Science and Technology Council (NSTC)

7.4.2 Open Research Products

Scientific knowledge, or research products, take the form of:

7.4.3 What is Data?

In general, data are pieces of information about a subject, including theoretical truths, raw measurements, or highly processed values.

There can even be data about data, called metadata. In our lessons, when we talk about data we are referring to scientifically or technically relevant information that can be stored digitally and accessed electronically such as:

- Information produced by missions and experiments, including calibrations, coefficients, and documentation
- Information needed to validate scientific conclusions of peer-reviewed publications

Open data can have many characteristics, including rich and robust metadata and be made available in a range of formats. These characteristics are detailed more later in this module, and even further in the module on Open Data.

7.4.4 What is Code?

Many scientists write source code to produce software to analyze data or model observations. Code is a language that humans can type and understand. Software is often a collection of programs, data, and other information that a computer system uses to perform specific tasks. Scientists write and use many different types of software as part of their research.

General Purpose Software – Software produced for widespread use, not specialized scientific purposes. This encompasses both commercial software and open-source software.

Operational and Infrastructure Software – Software used by data centers and large information technology facilities to provide data services.

Libraries – No creative process is truly complete until it manifests a tangible reality. Whether your idea is an action or a physical creation, bringing it to life will likely involve the hard work of iteration, testing, and refinement.

Just be wary of perfectionism. Push yourself to share your creations with others. By maintaining an open stance, you'll be able to learn from their feedback. Consider their responses new material that you can draw from the next time you're embarking on a creative endeavor.

Modeling and Simulation Software – Software that either implements solutions to mathematical equations given input data and boundary conditions, or infers models from data.

Analysis Software – Software developed to manipulate measurements or model results to visualize or gain understanding.

Single-use Software – Software written for use in unique instances, such as making a plot for a paper, or manipulating data in a specific way.

Some of the tools that you can use to develop software are introduced in Lesson 4. Understanding how to find and use others' code, create your own, and share it are an important part of advancing science and covered in the module on Open Code.

7.4.5 What are Results?

Results capture the different research outputs of the scientific process. Publications are the most common type of results, but this can include a number of other types of products. Both data and software can be considered a type of result, but when we discuss results, we will focus on other types of results. Results can include the following:

- Peer-reviewed publications
- Computational notebooks
- Blog posts
- Videos and podcasts
- Social media posts
- Conference abstracts and presentations
- Forum discussions

You may already be familiar with the research life cycle, but still unfamiliar with the types of results that can be shared openly throughout this process. When sharing results, we strive to be as open as possible, with the goal of increasing reproducibility, accessibility, and inclusion of our science. Throughout the research lifecycle, there are multiple opportunities to openly share different results that can lead to new collaborations and lines of inquiry. Additional details on the scope of open results are shared in Module 5 – Open Results.

7.5 Using Tools for Open Science in Practice

The following lessons in this module explore different tools and resources available to researchers for using, making, and sharing open science. As mentioned, it is important to

think about how to integrate open science principles across all stages of the research process. Here is an overview of one way the various pieces might work together.

7.5.1 The Components of Open Science

The four principal components of open science can be organized in a pyramid of openly-shared research products.

The research paper, closely tied to the results, sits at the top of the pyramid and summarizes how you've combined your software and your data to produce your results.

The practice of sharing these components can occur at varying degrees of completeness. For the following guidance on how to share components of open science, we simplify the range of completeness to "good", "better", and "best." This range reflects one's commitment to sharing open science at all steps in the research process and to all of its products.

7.5.2 Sharing Open Data

Data can be easily shared through many different services - the best way for scientific data to be shared is often through a long term data repository that will both preserve your data and make it discoverable. The image provides some of the considerations when sharing the data through Zenodo, a generalist data repository. These considerations would be similar for other data repositories. See Module 3 - Open Data for more details on sharing open data.

7.5.3 Sharing Open Code

When sharing open code, it is often through an online version controlled platform that allows others to contribute to the software and provides a history of changes to the software. For example, many researchers choose to post code files on GitHub with a BSD 3-Clause license. This permits others to contribute and reuse the software. Steps to preserve code and make it discoverable are discussed in Module 4 - Open Code.

7.5.4 Sharing an Open Paper

Researchers can choose to publish in a journal with an open access license. Researchers can search for open access journals through the Directory of Open Access Journals (DOAJ). (See Module 5 - Open Results)

7.5.5 Sharing Open Results

When sharing results, include your methodology that was used to produce results (i.e. the "provenance") directly with your software. Software tends to evolve with time while the outputs of the software itself can retain some consistency. Therefore, sharing your methodology helps others to reproduce your aging results with newer software, even if the methodology to produce them can vary as the software evolves.

7.5.6 An Open Science Project Example

Here is an example of how one group openly shared their data, results, software, and paper; all with their own unique identifiers. Note that data and software can each have multiple identifiers, enabling others to cite all versions or one unique version.

Here, you can review the separate elements of the image above. Select "<" and ">" buttons to navigate.

Data

This version: https://doi.org/10.5281/zenodo.3688691

All versions: https://doi.org/10.5281/zenodo.3688690

Results

https://doi.org/10.1175/JHM-D-19-0084.1

Software

This version: https://github.com/c-h-david/rapid

All versions: https://doi.org/10.5281/zenod

7.6 Lesson 1: Summary

In this lesson, you learned:

- Scientific knowledge, or research products, take the form of: data, software, and results.
- In general, data are pieces of information about a subject, including theoretical truths, raw measurements, or highly processed values.

7.7 Lesson 1: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/03

Read the statement below and decide whether it's true or false:

Open Science is the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility, and equity.

- True
- False

Question

02/03

What are the four principal outputs of open science? Select all that apply.

- Budgets
- Results
- Inclusion
- Data
- Code
- Paper

Question

03/03

Which of the following is a type of software? Select all that apply.

- General purpose
- Operational and Infrastructure
- Modeling and Simulation
- Analysis
- Material

8 Lesson 2: General Tools for Open Science

8.1 Navigation

8.2 Overview

This lesson introduces you to the commonly used tools in open science. It starts out by providing a brief introduction to open science tools and describes persistent identifiers - one of the most common open science tools in use that ensures reproducibility, accessibility, and recognition of scientific products. This is followed by descriptions of other common open science tools that are applicable regardless of your field of study. The lesson wraps up with a description of open science and data management plans that is a key component to sharing your science throughout the research process.

8.3 Learning Objectives

After completing this lesson, you should be able to:

- Recall the definition of open science tools.
- Describe what a persistent identifier is and state an example.
- List a few commonly used open science tools that support research.
- List the components of an Open Science and Data Management Plan and what they include.

8.4 Introduction to Open Science Tools

The word "tools" refers to any type of resource or instrument that can be used to support your research. In this sense, tools can be a collection of useful resources that you might consult during your research, software that you could use to create and manage your data, or even human infrastructure such as a community network that you join to get more guidance and support on specific matters.

In this context, open science tools are any tools that enable and facilitate openness in research, and support responsible open science practices. It is important to note that open science tools are often open source and/or free to use, but not always.

Open science tools can be used for: - **Discovery** - Tools for finding content to use in your research. - **Analysis** - Tools to process your research output, e.g. tools for data analysis and visualization. - **Writing** - Tools to produce content, such as Data Management Plans, presentations, and preprints. - **Publications** - Tools to use for sharing and/or archiving research. - **Outreach** - Tools to promote your research.

In this lesson, we introduce you to some of the most general open science tools such as persistent identifiers, metadata, documentation, and open science and data management plans. Regardless of the field of study, these tools and practices are some of the things that you will encounter as you use, make, or share your research. Read more about open science tools on OpenSciency.

8.5 Persistent Identifiers

A digital persistent identifier (or "PID") is a "long-lasting reference to a digital resource" that is machine-readable and uniquely points to a digital entity, according to ORCID examples of persistent identifiers used in science are described below.

8.5.1 ORCID

An "Open Researcher and Contributor Identifier" (ORCID) provides valid information about a person. Following are some key details about ORCIDs.

A free, nonproprietary numeric code that is:

- Uniquely and persistently identifies authors and contributors of scholarly communication.
- Similar to tax ID numbers for tax purposes.

ORCIDs are used to link Used to link researchers to their research and research-related outputs. It is a 16-digit number that uniquely identifies researchers and is integrated with certain organizations (like some publishers) that will add research products (such as a published paper) to an individual's ORCID profile. ORCIDs are meant to last throughout ones career, and helps to avoid confusion when information about a researcher changes over time (e.g. career change or name change). (cite: https://orcid.org/)

Many publishers, academic institutes, and government bodies support ORCID. In 2023, ORCID reported over 1,300 member organizations and over 9 million yearly live accounts. You can connect it with your professional information (affiliations, grants, publications, peer review, and more).

8.5.2 Digital Object Identifiers (DOI)

A DOI is a persistent identifier used to cite data, software, journal articles, and other types of media (including presentation slides, blog posts, videos, logos, etc.).

Unlike dynamic transient URLs, DOIs are static pointers to documents on the internet. Since a DOI is static, each new version of data or software that you want to cite will need a new DOI. Some DOI providers allow for one DOI to point to "all versions" and a series of individual DOIs for each specific version. Individuals cannot typically request a DOI themselves, but rather have to go through an authorized organization that can submit the request.

Making a DOI for your product ensures its longevity! This means, if you cite a DOI in a research paper, you can be confident that future readers will be able to follow that citation to its source, even if websites have completely changed in the meantime.

For example, the DOI: 10.5067/TERRA-AQUA/CERES/EBAF-TOA_L3B004.1 will always resolve to a web page that explains what the CERES_EBAF-TOA_Edition4.1 data set is and how to download it. (See the screenshot below if you're curious what this dataset actually is!)

DOIs are provided and maintained by the International Organization for Standardization (ISO): https://www.doi.org/.

8.5.3 Citations Using DOIs

DOIs make citing research products easier and more useful.

Data repositories will typically instruct you on the exact way to cite their data, which includes the correct DOI. For example, let's take a look at the CERES_EBAF-TOA_Edition4.1 data set mentioned above. This is an example from the Atmospheric Science Data Center's (ASDC) website.

8.5.4 Activity 2.1: Find and Resolve a DOI

In this activity, you will search for a DOI for a data set or piece of software that you use, and you will then use the DOI website to "resolve" the DOI name. By "resolving", this means that you will be taken to the information about the product designated by that particular DOI.

1. Find the DOI for a dataset or software you use often.
    1. This should be listed either in the citation file, or in the website where that data/software is published.
    2. If you can't find a DOI, you can instead locate the DOI listed on this page: https://asdc.larc.nasa.gov/project/CERES/CERES_EBAF-TOA_Edition4.1
2. Go to https://www.doi.org/ and scroll down to the bottom of the page to "TRY RESOLVING A DOI NAME".
3. Copy and paste the DOI you found into the form called "TRY RESOLVING A DOI NAME".
4. Click Submit.
5. The page should automatically redirect you to a page that explains and contains the cited data.

Activity Takeaways: Find and Resolve a DOI

This activity will vary depending on which DOI you choose to use. However, if you used the example presented, you should find the DOI: 10.5067/TERRA-AQUA/CERES/EBAF-TOA_L3B004.1

And after step 5, you should end up back on the page https://asdc.larc.nasa.gov/project/CERES/CERES_EBAF-TOA_Edition4.1

This is how easy it should be for your readers to find and use your citation information.

8.5.5 Examples of PIDs in Action

Example 1

Example 2

Example 3

The necessity for a persistent identifier (PID) begins when a researcher writes code. To make the code searchable, the researcher uploads their code to a repository and registers a DOI for their script. Now others can review and use the code, and cite it properly.

Example 1

Example 2

Example 3

A workshop planning committee collaboratively authors a paper that summarizes the results of a workshop. They collect the ORCIDs of everyone who participated in the workshop, and include them in the paper. Finally, they publish in an academic journal that automatically assigns the paper a DOI.

Example 1

Example 2

Example 3

A community scientist attends an online conference and gives a short talk. They deposit their slides in an online repository, then create a DOI to enable easy sharing with colleagues and straightforward citation.

8.6 Useful Open Science Tools

8.6.1 Metadata

Metadata are data that describe your data, either accompanying your data as a separate file or embedded in your data file. They are often used to provide a standard set of general information about a dataset (e.g., data temporal/spatial coverage or data provider information) to enable easy use and interpretation of the data.

Metadata is essential to the implementation of FAIR Principles because it makes data searchable in an archive, provides context for future use, and presents a standard vocabulary.

Metadata can be more readily shared than data - it usually does not contain restricted information and it is much smaller than the entire data set.

8.6.2 Purpose of Metadata

Metadata can facilitate the assessment of dataset quality and data sharing by answering key questions, such as information about:

- How data were collected and processed.
- What variables/parameters are included in the dataset.
- What variables are and what variables are related to.
- Who collected the data (science team, organization, etc.).
- How and where to find the data (e.g., DOI).
- How to cite the data.
- Which spatio-temporal region/time the data covers.

- Any legal, guideline, or standard information about the data.

Metadata enhances searchability and findability of the data by potentially allowing other machines to read and interpret datasets.

According to The University of Pittsburgh, "A metadata standard is a high level document which establishes a common way of structuring and understanding data, and includes principles and implementation issues for utilizing the standard."

Many standards exist for metadata fields and structures to describe general data information. It is a best practice to use a standard that is commonly used in your domain, when applicable, or that is requested by your data repository. Examples of metadata standards for different domains include:

- CF Metadata Conventions
- World Meteorological Organization WIS 2.0
- GeneLab Working Group

8.6.3 Types of Metadata

There are different types/categories of metadata addressing different purposes:

Descriptive Metadata

Structural Metadata

Administrative Metadata

Descriptive metadata can contain information about the context and content of your data, such as variable definition, data limitation, measurement/ sampling description, abstract, title, and subject keywords.

Descriptive Metadata

Structural Metadata

Administrative Metadata

Structural metadata are used to describe the structure of the data (e.g., file format, the dataset hierarchy, and dimensions).

Descriptive Metadata

Structural Metadata

Administrative Metadata

Administrative metadata explains the information used to manage the data (e.g., when and how it was created, which software and the version of the software used in data creation).

8.6.4 Documentation

Documenting the production and management of your science benefits both you and those that might use your data, code, or results in the future. You are your own best collaborator. Documentation can save you from a headache should you need to reference or reuse your work in six months or attempt to recall meticulous details about your process later on. Properly documented research products increase their usability.

Types of documentation include (many of which will be expanded upon later in this curriculum):

Data

Software

Results

Summary of the data (e.g., as a README file or user guide) that answers questions such as:

What are known errors for these data?

How can this data be used?

How were the data collected?

Associated publications – how did others use these data?

Data

Software

Results

README files: Basic installation and usage instructions.

Inline comments in code: Annotations on code components.

Release notes: What is new in this version?

Associated publications: How did others use this software?

Data

Software

Results

Associated publications: What was the research process?

Packages of data and software for regenerating results.

8.6.5 Repositories

Repositories are storage locations for data, results, code and compiled software, providing the most common way to share and find each of these components. In general, you want to use a long term repository that will independently host and store your data making sure that it is both shared and preserved. Different kinds of repositories serve different purposes. For example, Zenodo acts as an archiving repository for individual version releases of data, software, and publications.

Different types of repositories:

- General repositories
- Domain-specific repositories
- Institutional repositories
- National repositories

Users should select repositories based on their needs. See the lessons in the rest of this module and Modules 3-5 for more details.

8.6.6 Pre-registration

Pre-registration is the process by which a researcher documents their research plans in an open access format prior to the start of a project. This provides a locked, time-stamped proof of the origin of a concept. Pre-registration is currently more widely adopted by certain disciplines, particularly the social sciences.

Types of Pre-Registration Include:

Standard Pre-registration

Registered Reports

Registered Replication Report

Sharing Grant Proposals

An investigator documents their plans in writing and submits them to a pre-registration service. This documents the researcher's plans prior to undertaking the research, and provides investigators and reviewers with a way to distinguish a priori hypotheses from post-hoc exploratory analyses. The document may be kept private for some period of time, but is usually made public upon submission of the manuscript for publication.

Standard Pre-registration

Registered Reports

Registered Replication Report

Sharing Grant Proposals

An investigator writes a manuscript describing the motivation for a study and a detailed description of the methods, and submits it to a journal for peer review prior to undertaking the research. The manuscript is reviewed based on the importance of the research question and the quality of the methods. If accepted, the journal agrees to publish the paper regardless of the results, assuming that there are no problems with the implementation of the methods.

Standard Pre-registration

Registered Reports

Registered Replication Report

Sharing Grant Proposals

A type of registered report in which the investigators wish to attempt to replicate a particular published finding, usually involving multiple research sites.

Standard Pre-registration

Registered Reports

Registered Replication Report

Sharing Grant Proposals

Another way to document and timestamp research plans and concepts is to share funded grant proposals publicly. This has the added benefit of making the funding process more transparent, and providing examples of successful grant proposals for other researchers, particularly those in their early career stage.

8.6.7 Why is Pre-Registration Important?

- It forces the researcher to plan and think through both why and how they are pursuing their research question.
- It provides the researcher with a way to determine whether a hypothesis was truly held a priori, versus relying upon memory.
- It forces the researcher to think through their analysis plan in more detail, potentially surfacing issues that could influence the design of the study.
- It helps prevent unethical manipulation of data analyses and project design to yield statistically relevant results.
- Helps prevent selective reporting of measures.

8.6.8 When Can/Should One Pre-Register Their Research?

A planned research activity can be pre-registered at any point, as long as the particular activity being registered has not started. However, there are several points at which registration is most common:

- Prior to the collection of data for a project
- Prior to analysis of an existing or openly available dataset

Source: Registration — Stanford Psychology Guide to Doing Open Science (pol-drack.github.io)

A 2023 Nature survey on researcher attitudes towards open science practices found that about 88% of respondents favor sharing data or code online while only 58% support pre-registration. This moderate support for pre-registration among respondents suggests that awareness of its benefits and lingering concerns remain issues. In the next section, we introduce a method to strategize how to best implement open science from the beginning of a study to its end.

8.7 Open Science and Data Management Plans

To successfully use, make, and share science openly, we need an open science and data management plan (OSDMP).

- From day 1, establish a plan for management, preservation, and release of data, software, and results.
- This plan is your blueprint for open science - refer to your plan often to ensure you succeed in your goal of openness.

We'll discuss each component (data, software, & results) when we cover each topic.

8.7.1 Design Your Science to be Open

Funding organizations and agencies around the world are beginning to require open science plans.

The OSDMP describes how the scientific information that will be produced from scientific activities will be managed and made openly available. Specifically, a plan should include sections on data management, software management, and publication sharing. If your study has other types of outputs, such as physical samples, hardware, or anything else, you should include those in the plan. An OSDMP helps researchers think about the details of how they plan to share results.

A well written OSDMP can help you win funding because it demonstrates your skills at doing open science!

Example sections to include in an OSDMP:

1. Data Management Plan (DMP)
2. Software Management Plan (SMP)
3. Publication sharing
4. Other open science activities
5. Roles and responsibilities

The steps for each of these sections should include:

- What?
  - Description of types of materials that will be produced
- When?
  - The schedule for archiving and sharing
- Where?
  - The repository(ies) and archives that will be used to share materials
- How?
  - The details of how to enable reuse of materials (eg. licensing, documentation, metadata)
- Who?
  - Roles and responsibilities of the team members

8.7.2 Data Management Plan

Every major research foundation and federal government agency now requires scientists to file a data management plan (DMP) along with their proposed research plan. Data and other elements such as code and publications have their own lifecycle and workflow, which need to be in the plan. DMPs are a critical aspect of open science and help keep other researchers informed and on track throughout the data management lifecycle.

DMPs that are successful typically include a clear terminology about FAIR and CARE principles and how they will be applied.

The data management lifecycle is typically circular. Research data are valuable and reusable long after the project's financial support ends. Data reuse can extend beyond our own lifetimes. Therefore, when designing a project or supporting an existing corpus of data, we need to remain cognizant of what happens to the data after our own research interaction ends.

Data management plans typically include the following:

- Descriptions of the data expected to be produced from the proposed activities, including types of data to be produced, the approximate amount of each data type expected, the machine-readable format of the data, data file format, and any applicable standards for the data or associated metadata.
- The repository (or repositories) that will be used to archive data and metadata arising from the activities and the schedule for making data publicly available.
- Description of data types that are subject to relevant laws, regulations, or policies that exclude them from data sharing requirements.
- Roles and responsibilities of project personnel who will ensure implementation of the data management plans.

8.7.3 Software Management Plan

Software management plans describe how software will be managed, preserved, and released as part of the scientific process. This helps ensure transparency and reproducibility in the scientific process. Module 4 on Open Code shares more details about the importance of sharing code as part of the scientific process.

General components of a software management plan:

- Description of the software.
- Repository(ies) and archive(s) in which software will be shared.
- Sharing guidelines.
- Personnel roles and responsibilities.
- Any community-specific information of note.

At a minimum, a software management plan for SMD-funded research should include: - Description of the software expected to be produced from the proposed activities, including types of software to be produced, how the software will be developed, and the addition of new features or updates to existing software. This can include the platforms used for development, project management, and community-based best practices to be included such as documentation, testing, dependencies, and versioning. - The repository(ies) that will be used to archive software arising from the activities and the schedule for making the software publicly available. - Description of software that are subject to relevant laws, regulations, or policies that exclude them from software sharing requirements. - Roles and responsibilities of project personnel who will ensure implementation of the software management plan.

8.7.4 Open Science Plan

The OSDMP should also describe other open processes as part of the plan. This includes the types of publications that are expected to be produced from the activities, including peer

reviewed manuscripts, technical reports, conference materials, and books. The plan should also outline the methods expected to be used to make the publications publicly accessible.

This section may also include a description of additional open science activities associated with the project. This may include:

- Holding scientific workshops and meetings openly to enable broad participation.
- Pre-registering research plans in advance of conducting scientific activities.
- Providing project personnel with open science training or enablement (if not described elsewhere in a proposal).
- Implementing practices that support the inclusion of broad, diverse communities in the scientific process as close to the start of research activities as possible (if not described elsewhere in a proposal).
- Integrating open science practices into citizen science activities.
- Contributions to or involvement in open-science communities.

8.7.5 Publications Plan

A plan for publications is a crucial piece of the OSDMP. A publications plan should include the following features:

- Describes how results will be managed, preserved, and released - in other words, how you will communicate your findings.
- Includes plans for conference talks, whitepapers, peer reviews journal articles, books, and other such documents.
- Written in compliance with any rules and regulations within your organization, as well as from your funding source.
- As with the data and software plans, it serves as a foundational framework for your project from start to finish.

8.7.6 Examples of Requirements for Open Science Management Plans

Globally, organizations and agencies are moving towards open science and beginning to require plans as part of funding. Here are just some of them:

USA - **NASA** - Open Science and Data Management Plan - **NSF** - Data Management Plan - **NIH** - Data Management and Sharing Plan - **NOAA** - Data and Information Sharing Plan (DISP)

GLOBAL INSTITUTES - **Australian Research Council** - Data Management Plan - **EU Open Science Requirements** - https://openscience.eu/Open-Science-in-Horizon-Europe -

UK Wellcome Trust - (Output Management Plan) - **Korea's National Research Foundation (NRF)** - DMP Guideline - **Japan Science & Technology Agency (JST)** - Open Access to Research Publications and Research Data Management

And remember, open science is nuanced! Although one of the tenants of open science is to share your products, not all products can or should be shared. How you share them may be specified by your organization or funding agency. As you embark on adopting open science for a project, consider if the subject and approach to your project will allow for sharing. Think about the following questions:

- Can the research products be shared?
- Who helped you obtain your data?
- Will they benefit from release?
- Who has responsibility and/authority for what happens with the data?
- Should the research products be shared?

More details on how to write these plans for data, code, and results are in the following modules.

8.8 Lesson 2: Summary

In this lesson, we learned:

- The definition of science tools, common examples, and which part of the scientific workflow they can support.
- The definition and purpose of persistent identifiers. The usefulness of ORCIDs and DOIs in the scientific process.
- Examples of useful and common open science tools such as metadata, documentation, repositories, and pre-registration.
- The steps for writing an open science and data management plan.

8.9 Lesson 2: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/03

What can open science tools help with?

- Discovery
- Writing

- Outreach
- All of above

Question

02/03

Complete the statement:

Good, clear Metadata _____.

Select all that apply.

- Improves findability
- Improves accessibility
- Improves interoperability
- Improves reusability
- Is a waste of time

Question

03/03

Which are the components of a Software Management Plan? Select all that apply.

- Description of the software
- Repository(ies) in which software will be archived
- Sharing guidelines
- Personnel roles and responsibilities
- Any community-specific information of note
- Creating a logo

9 Lesson 3: Tools for Open Data

9.1 Navigation

9.2 Overview

This lesson discusses the concepts, considerations, and tools for making data and results. It starts with a closer look at the FAIR principles and how they apply to data. The lesson includes an introduction to plans, tools, data formats, and other considerations that are related to making data and sharing the results related to that data.

9.3 Learning Objectives

After completing this lesson, you should be able to:

- Define the different types of scientific data.
- Define what the acronym FAIR means and explain how it supports the sharing of open data.
- Identify data management practices and tools to locate data in repositories.
- List and explain the purpose of the resources commonly used in making data including the data formats, inspecting data, and assessing 'FAIR'-ness of data.

9.4 Introduction to Open Data

Data is a major part of scientific research, and why wouldn't it be? It informs tools that we use, stories that we read, and decisions that we make on a daily basis.

For instance, the open access Copernicus Emergency Management Service implemented by the European Commission produces 24/7 open access data collected by ESA and NASA satellites to produce maps that inform disaster preparedness and response efforts across the globe. This is only one example among many others demonstrating the value of data, particularly open and public data, in our daily life and for public good.

Data shared openly in scientific research brings tremendous value to the scientific community and beyond, from indigenous communities to urban populations. Before understanding the broad based impact of data, let's first look at what is data in the context of scientific research. Specifically, we will discuss the definition and characteristics of open data?

9.4.1 What is Data?

Scientific data is any type of information that is collected, observed, or created, in the context of research. It can be:

- Primary – Raw from measurements or instruments
- Secondary – Processed from secondary analysis and interpretations.
- Published – Final format available for use and reuse.
- Metadata – Data about your data.

It is everything that you need to validate or reproduce your research findings, as well as what is required for the understanding and handling of the data.

The following sections discuss ways to ensure that data is fully utilized and accessible to the most amount of people. These best practices center around community frameworks and tools that help researchers manage and share open data.

9.5 FAIR Principles

Just like driving on a road, if everyone follows agreed upon rules, everything goes much smoother. The rules don't need to be exactly the same for every region, but share common practices based on insights about safety and efficiency.

For example, maybe you drive on the left side of the road or the right side. Either is fine, those sort of details are for different communities to decide on. However, there are overarching guidelines shared by communities across the globe, such as the rule to drive on the road not the sidewalk, use a turn signal when appropriate, adhere to lights at intersections that direct

traffic, and follow speed limits. Some communities may implement stricter rules than others, or practice them differently, but these guidelines help everyone move around safely through a common understanding of how to drive on roads. For scientific data, these guidelines are called the Findable, Accessible, Interoperable, Reusable or "FAIR" principles. They do to data what their title suggests. That is, these principles make it possible for others (and yourself) to find, get , understand, and use data correctly.

Findable:

To be Findable:

- Data and results are assigned a globally unique and persistent identifier.
- Data are described with rich metadata.
- Metadata clearly and explicitly include the identifier of the data it describes.
- Data and results are registered or indexed in a searchable resource.

Current Enabling Tech: - DataCite's Metadata Schema - PIDs: Persistent IDentifiers (additional details in the following sections) - Digital Object Identifier (DOI): A top-level and a mandatory field in the metadata of each record - for data, code, publications. - Open Research and Contributor ID (ORCiD) - A code that uniquely identifies authors and contributors of research products and scholarly communication.

Accessible

To be Accessible:

- Data and results are retrievable by their identifiers using a standardized communication protocol.
- The protocol is open, free, and universally implementable.
- The protocol allows for an authentication and authorization procedure, where necessary. Data and results are publicly accessible and licensed under the public domain.
  - Metadata are accessible, even when the data are no longer available Data and metadata will be retained for the lifetime of the repository.
  - Metadata are stored in high-availability database servers.

Current Enabling Tech: - File Transfer Protocol (FTP), File Transfer Protocol Secure (FTPS) - Hypertext Transfer Protocol (HTTP), Hypertext Transfer Protocol Secure (HTTPS)

Note that Microsoft Exchange Server and Skype are examples of proprietary protocols.

Interoperable

To be Interoperable:

- Data uses a formal, accessible, shared, and broadly applicable language for knowledge representation.
- Data uses a known, standardized data format.

- Data use vocabularies that follow FAIR principles.
- Data include qualified references to other (meta)data.

Current Enabling Tech:

- Zenodo uses JSON Schema as internal representation of metadata and offers export to other popular formats such as Dublin Core or MARCXML.
- For certain terms we refer to open, external vocabularies, e.g.: license (Open Definition), funders (FundRef) and grants (OpenAIRE).
- Each referenced external piece of data is qualified by a resolvable URL.

Reusable

To be Reusable:

- Data are richly described with a plurality of accurate and relevant attributes.
  - Data are released with a clear and accessible data usage license.
  - Data are associated with detailed provenance.
  - Data meet domain-relevant community standards.

Current Enabling Tech: - The metadata record contains a minimum of DataCite's mandatory terms, with optionally additional DataCite recommended terms and Zenodo's enrichments. - Zenodo is not a domain-specific repository, yet through compliance with DataCite's Metadata Schema, metadata meets one of the broadest cross-domain standards available.

Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci.Data 3:160018, doi: 1 0 .1038/sdata.2016.18 (2016)

These are high-level guidelines, and much like open science, implementation is nuanced. Sometimes it takes a group effort and/or a long production process/funding to make data and results FAIR. For other datasets, it could be more straightforward. A well-coordinated data management plan is needed for full compliance with FAIR, and the details of this will be discussed further in Module 3 – Open Data.

9.6 Tools to Help with Planning For Open Data Creation

9.6.1 Data Management Plan

The previous lesson describes the requirements of a data management plan (DMP). Below are two open science resources to get you started or creating a data management plan:

DMPTool

The DMPTool in the US helps researchers by featuring a template that lists a funder's requirements for specific directorate requests for proposals (RFP). The DMPTool also publishes other

open DMPs from funded projects that can be referenced to improve your own. The Research Data Management Organizer (RDMO) enables German institutions as well as researchers to plan and carry out their management of research data.

ARGOS

ARGOS is used to plan Research Data Management activities of European and nationally funded projects (e.g. Horizon Europe, CHIST-ERA, the Portuguese Foundation for Science and Technology - FCT). ARGOS produces and publishes FAIR and machine actionable DMPs that contain links to other outputs, e.g. publications-data-software, and minimizes the effort to create DMPs from scratch by introducing automations in the writing process. OpenAIRE provides a guide on how to create DMP.

9.6.2 Data Repositories

A data repository is a digital space to house, curate, and share research outputs. Data repositories were originally used to support the needs of research communities. Examples of data repositories include:

- **Protein Data Bank** uses a data repository to catalog 3D structures of proteins and nucleic acids.
- **Genbank** of the National Institutes of Health uses a genetic sequence database that contains annotated publicly available nucleic acid sequences.
- **The Image Data Resource** is a public repository of microscopy bio-image datasets from published studies.
- **The Electron Microscopy Public Image Archive** is a public resource for raw cryo-EM images.
- **OpenNeuro** is an open platform for validating and sharing brain imaging data. The tools featured in Open Neuro enable easy access, search, and analysis of annotated datasets.

Open science tools such as data repositories should implement FAIR principles, especially in the case of attribution of persistent identifiers (e.g., DOI), metadata annotation, and machine-readability.

ZENODO

Zenodo is an example of a data repository that allows the upload of research data and creates DOIs. Its popularity among the research community is due to its simplified interface, support of community curation, and feature that enables researchers to deposit diverse types of research outputs; from datasets and reports to publications, software, multimedia content.

DATAVERSE

The Dataverse Project is an open source online application to share, preserve, cite, explore, and analyze research data, available to researchers of all disciplines worldwide for free.

DRYAD

The Dryad Digital Repository is a curated online resource that makes research data discoverable, freely reusable, and citable. Unlike previously mentioned tools, it operates on a membership scheme for organizations such as research institutions and publishers.

DATACITE

Datacite is a global non-profit organization that provides DOIs for research data and other research outputs, on a membership basis.

OSF

The Open Science Framework is an open source platform for sharing, managing, and collaborating research.

Data services and resources for supporting research require robust infrastructure which relies on collaboration. An example of an initiative on the infrastructures of data services comes from the EUDAT Collaborative Data Infrastructure, a sustained network of more than 20 European research organizations.

Private companies also host and maintain online tools for sharing research data and files. For example, Figshare is one example of a free and open access service operated by private companies. It provides DOIs for all types of files and recently developed a restricted publishing model to accommodate intellectual property (IP) rights requirements. It allows sharing the outputs only within a customized Figshare group (could be your research team) or with users in a specific IP range. Additional advances include integration with code repositories, such as GitHub, GitLab, and Bitbucket.

Additional research data repositories can be found in the publicly available Registry of Research Data Repositories. OpenAire, a hosted search engine, also provides a powerful search function of data and repositories. It features a filter for country, type, and thematic area, as well as enables the download of data.

The amount of data, repositories, and different policies can be overwhelming. When in doubt of determining which repository is right for you, consult librarians, data managers and/or data stewards in your institution, or check within your discipline-specific or other community of practice.

9.6.3 Activity 3.1: Explore Zenodo and Sign Up!

Explore open repositories to familiarize yourself with their structure and available product information. The most popular repository at the moment is Zenodo. Review the following 4.5-minute video to get an overview of Zenodo and then sign up for an account. You can use your ORCID to sign up if you have one or made one in the previous lesson.

Watch Video

9.7 Tools to Help with Using and Making Open Data

9.7.1 Data Formats

A useful file format can be read into memory by some software. Think of the format as a tool for making data accessible. Easy to use formats feature:

- A simple, easy to understand structure.
- A clear and open specification for the format that is ideally not tied to a specific software product.
- Open software libraries and APIs that can parse the format.

The formats that are considered the most interoperable against the criteria above include Comma Separated Values (CSV), Extensible Markup Language (XML), and JavaScript Object Notation (JSON). Other common formats for researchers include binary array-based formats like Network Common Data Form (NetCDF), Hierarchical Data Format (HDF), Geotiff, Flexible Image Transport System (FITS), and other formats designed for cloud storage and access like Zarr, Cloud Optimized GeoTIFF, and Parquet. Many of these formats have tools that check datasets for compliance and readability.

9.7.2 Inspecting Data

Modern data formats allow the storage of much more than mere data points. Once one adopts these standards (e.g. NetCDF), the discovery of the contents on each file can be aided by a variety of tools which together help map primary data and/or display the associated metadata. Several tools exist for inspecting data, too numerous for all to be mentioned here. Notable tools to start with include:

CSV, XML, JSON - These files can all be opened with most common text editors. There are some tools that can create views of the files that are more user-friendly, such as: - csv: Microsoft Excel and Google Sheets - xml: Most internet browsers and with any text editor like Notepad or Microsoft Word or Google Docs - json: http://json.parser.online.fr/ and https://jsonformatter.org/json-pretty-print

NetCDF, HDF, FITS - These files require special software tools to view their contents. Many of these tools will also visualize the data as well. - NetCDF and HDF: Most files are easily viewed using the Xarray open-source software library in Python or the ncdf4 library in R. - FITS: There are many options, a list is provided at https://fits.gsfc.nasa.gov/fits_viewer.html

ZARR, COG, PARQUET - These files require special software tools to view their contents. Many of these tools will also visualize the data as well. - Zarr: Files are easily viewed using the Xarray open-source software library in Python or the Pizzar library in R. - COG: Files are viewed using the rioXarray open-source software library in Python or the terra library in R. - Parquet: Files are viewed using the Pandas open-source software library in Python or the Arrow library in R.

9.7.3 FAIR Assessment

How 'FAIR' is your data? Two groups - FAIRsharing.org and the Research Data Alliance (RDA) - have developed the FAIR Metrics and FAIR Data Maturity Model to help assess the 'FAIR'-ness of a dataset. There are open-source tools that help researchers assess their data:

AUSTRALIAN RESEARCH DATA COMMONS (ARDC)

Online questionnaire (manual) Best for:

- Triggering discussions at the initial stages of considering FAIR implementation
- Identifying areas for improvement

Outputs include:

- Progress bar for each FAIR principle
- Aggregate bar for all

FAIR-CHECKER

Automated via website or API

Best for:

- Scalability to many datasets
- Identifying areas for improvement

Outputs include: - A chart with scores and details

F-UJI

Automated via website or API

Best for:

- Scalability to many datasets
- Detailed documentation of tool

Outputs include: - A report and chart with scores and details

FAIR EVALUATION SERVICES

Automated via website or API

Best for:

- Scalability to many datasets
- To generate a custom assessment

Outputs include:

- A detailed report and chart

9.8 Lesson 3: Summary

In this lesson you learned:

- The different types of scientific data, including primary, secondary, published, and metadata.
- A list of open science practices to implement FAIR principles that make data and results easily accessible to a wide range of people.
- Digital tools to help plan for making and sharing open data.

9.9 Lesson 3: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/03

Choose the FAIR Principles from the list below. Select all that apply.

- Reproducible
- Reusable
- Responsible
- Findable
- Interactive
- Interoperable
- Interspersed

- Accessible
- Authorizable

Question

02/03

Which of the following can help make your data FAIR? Select all that apply. - Get a license for your data - Make sure you develop your own metadata - Obtain a PID for your data

Question

03/03

Which of the following are examples of repositories? Select all that apply.

- Zenodo
- Dataverse
- Dryad
- Datacite
- Google

10 Lesson 4: Tools for Open Code

10.1 Navigation

10.2 Overview

This lesson introduces you to some useful tools for working with open code. You will learn the various tools available to develop, store, and share open code, from version control to code editing software to containers.

10.3 Learning Objectives

After completing this lesson, you should be able to:

- Explain the benefits of using tools for open code development.
- Define version control and understand how it supports collaboration in the development and management of code.
- List a few tools for editing software and some of their features.
- Distinguish between software repositories and software archives.

10.4 Introduction to Open Code

In Lesson 3, we learned about useful tools for working with scientific data. Now, we will provide an overview of commonly used tools that help us write and run computer code to explore, analyze, and visualize our scientific data. Later in Module 4 – Open Code, we will

discuss in greater detail what it means to make our code open, and walk through the steps of how to find, create, and share open code.

Understanding how to work with scientific code is essential in the modern landscape of data-driven research. The tools presented in this lesson encompass a diverse array of resources designed to streamline, enhance, and optimize the process of developing, maintaining, and collaborating on code development for scientific research. They enable the creation of robust and efficient code, often leveraging the collective wisdom of the open-source community. In the pursuit of reproducibility and transparency, these tools can also facilitate the sharing and dissemination of scientific code, fostering collaboration and ensuring that the foundations of scientific research remain open and accessible to all.

10.4.1 Historical Precedent for Making Code Open: Linux Operating System

Is the idea of writing code openly a new concept? No!

10.4.1.1 Context: Development of Linux OS

CLICK TO LEARN

- Started in 1991 by Linus Torvalds.
- Almost immediately released for scrutiny.
- Many eyes → Many bugs found → Many fixes.

10.5 Tools for Version Control

10.5.1 Version Control

Version control is the practice of tracking and managing changes made to code or other types of files. You may be familiar with "Track changes" in software like Microsoft Word. This is a form of version control, though not one well-suited to working with code. Version control is considered standard practice in the software development community, and simplifies management of code through time.

The general way we use version control starts by initializing a folder on your computing platform with the version control system you are using. A version control system automatically tracks all changes made by contributors and allows you to work offline and return later with updates. You write code as you usually do in your code editor of choice. After you have written some code or made some updates to existing code, you then commit those changes to the version control system to create a sort of "checkpoint" that you can then revert back to later if necessary. Then you add or update more code, and commit changes again. Each

commit requires you to add a short message which lets you briefly describe what changes were made. These messages serve as metadata that ensures collaborators, future users, and future you understand your development process at a point in time.

This may sound like a simple process, and in many ways it is! So why is it so important? Especially when it comes to coding, the ability to create a snapshot in time of a piece of code can be very helpful. For instance, you may have a piece of code that yields the intended result, but then you want to add a new function. You may choose to copy that code file so you don't lose the current state, and then work in a new file. This can become cumbersome pretty quickly when you have multiple files that are different versions of the same piece of code. Or instead of creating a new file, you may write code for the new function directly in the original file, but now the code throws errors when you try to run it, and you can't remember which lines you added since the last time the code ran without errors. By using version control, these problems are solved because we can revert back to the checkpoint when the code ran cleanly, and thereby avoid the need to create multiple copies to save the original piece of code.

There are many other features of version control systems, such as the concept of creating "branches" that allow you to work on new updates to a piece of code independently from and in parallel to the original piece of code. A branch is a deviation from the original code, but can be merged back into the original code when desired. All of these concepts are even more useful when collaborating with others using version control platforms, a collaborative practice that will be discussed later in this lesson.

10.5.2 Types of Software Version Control

There are two main styles of software version control systems:

CENTRALIZED

DISTRIBUTED (MORE POPULAR)

Singular "main" copy of the codebase

Must interact with specific server

Example: Subversion (SVN)

CENTRALIZED

DISTRIBUTED (MORE POPULAR)

Control:

Each developer's system can retain a copy of the codebase Examples:

Git

Mercurial

Using a distributed version control system like Git gives you more flexibility.

Example: Git

The most popular version control system for software development is Git. Git is open-source and is commonly used in conjunction with web-based software hosting sites like GitHub and GitLab (more on these in the next section), which allow for collaboration and sharing of code. You can also use it on your local computer when writing your own code. Git is often run at the command line, but there are other interfaces for using Git as well, including GitHub Desktop and some code editors that have Git integration included (more on this later).

https://xkcd.com/1597/

Git is very powerful and widely used (according to a Stack Overflow developer survey, over 87% of developers use Git), but that doesn't mean it is straightforward to learn. There are many good resources for learning Git (see below). If you find Git confusing at first, know that you are not alone! (There's even an XKCD comic about it!). For in-depth training on Git, please see the Software Carpentry lesson, listed below: Version Control with Git: Summary and Setup (swcarpentry.github.io)

10.5.3 Version Control Platforms

Version control platforms, typically web-based software hosting platforms, expand the usefulness of version control by allowing for a centralized location to store and collaborate on code, along with many other helpful features for code development and sharing.

Some examples of version control platforms:

- GitHub: a Git-based platform that allows collaboration and code history tracking. Owned by Microsoft.
- GitLab: a Git-based platform that also offers DevOps and CI/CD functionalities.
- BitBucket: a platform that can host Git and Mercurial repositories. Owned by Atlassian.

GitHub is one of the most popular platforms, and so we will provide examples of how to use GitHub in the rest of this section. It is important to note that GitHub is where most open-source software packages are housed, and so if you are interested in getting more involved with the open source software community, GitHub is an essential tool to learn how to use!

Example: GitHub

GitHub is an online, cloud-based software repository hosting site that integrates with Git and offers many other features that help with code development, collaboration, testing, and

releases. Before we dive into some of these features, it's important to understand how GitHub acts as a remote repository when using version control systems like Git.

If we go back to the general idea of using version control systems, GitHub can be added into the picture as a remote repository that hosts code. After creating a "checkpoint" in Git, you can then upload a copy of the current snapshot of your code to GitHub. There are a few reasons you might want to do this, including:

- To serve as a backup for your work (it is now stored on a remote server that you can access even if your computer dies).
- To share your code with others (more on this later in this course).
- To collaborate with others on your code. By uploading to GitHub, your code can be made accessible to others who might want to add features.

Let's expand on some of GitHub's collaboration tools. Some of these features include:

Term

Description/Definition

Issue Tracking

Keep track of feature requests, bugs, and other types of updates via GitHub Issues. GitHub also allows the use of labels and assigning people to tasks to help organize tasks.

Project Discussion Forums

GitHub allows for an online discussion forum where you can ask and answer questions, and hold community discussions.

Contribution Tracking

GitHub has a straightforward way to keep track of suggested code contributions (called "Pull Requests") from different people.

Code Review Tools

GitHub has a rich set of tools for reviewing and accepting (or denying) contributions from others (or yourself), such as in-line comments and easily viewable tracked changes to individual files.

Tailored Permissions

Choose who has the ability to update the code. This helps you feel confident that only those with permission can update code that you shared in GitHub, and also others feel safe to suggest updates without worrying that they might accidentally overwrite existing code.

All of these features excel at enabling asynchronous collaboration across teams. Most scientific open- source packages use GitHub for their primary code development. Note that there are many more GitHub features that we don't go into here that support collaboration, as well as

automated workflows and so much more. To learn more about GitHub, take a look at these references:

- How to Use Git and GitHub – Introduction for Beginners (freecodecamp.org)
- Getting Started with GitHub — Pythia Foundations (projectpythia.org)

[Cite Project Pythia: https://foundations.projectpythia.org/preamble/how-to-cite.html]

10.5.4 Summary of Benefits to Using Version Control and Version Control Platforms

- Features the ability to rewind changes back to any committed point
- Eases collaboration with others
- Keeps a directory clean from clutter, with no need for multiple copies of files
- Provides a targeted backup system for your work

10.6 Tools for Editing Code

10.6.1 Integrated Development Environment (IDEs)

An Integrated Development Environment (IDE) plays an important role in open code development by offering a comprehensive toolkit to researchers, scientists, and developers for editing code. It is a software application that streamlines the entire process of creating, testing, and managing code for scientific research and data analysis. By providing an all-in-one platform, an IDE allows researchers to write, debug, and optimize code more efficiently, fostering collaboration and reproducibility in open code science projects.

In open science, where transparency and accessibility are paramount, IDEs often incorporate version control systems like Git to facilitate collaboration and ensure that a research codebase is readily available for others to use and improve. Additionally, many IDEs integrate with data analysis and visualization tools. This makes it easier for scientists to analyze and interpret their data, ultimately contributing to the advancement of open code science practices.

If you were in a room with 10 developers and asked them each what their favorite code editor is, you would get many different responses. In this lesson, we will go over a few of the more popular varieties.

Source-Code Editing & Kernels – The Value of IDEs and Kernels

IDEs can bring a lot of good tools to your efforts. It's not just about editing code any more. Modern, robust IDEs can do most of the things listed here, if not more. One can use an IDE without executing in a kernel; one can use a kernel without having developed code in an IDE. However, they can work hand-in-hand.

Integrated Development Environment (IDE)

Kernel

Source code editing:

Syntax highlighting

Error/bug warnings

Plugins

Debuggers

Memory management

Version control

Build automation

Integrated Development Environment (IDE)

Kernel

Execution environment

Like a virtual machine

Isolates work area

Tailor settings

Easily replicable

IDE Example: Visual Studio Code

The most popular IDE these days, Microsoft's Visual Studio Code (or VS Code) is feature-rich without being clunky.

- It has a "dark mode" option which is easier on the eyes for long coding sessions.
- It provides the basics such as syntax highlighting and an integrated terminal window.
- It also has a wealth of plugins for connecting to servers, version control systems, and troubleshooting. It has several linter plugins, which can analyze your code for bugs, errors, and to help your team code in a consistent "style". This eases code maintenance down the road.
- If your line of code has an obvious error in it, the IDE will produce a red squiggle, just as if you've spelled something wrong in a Word Document.

Below is an example of a developer who accidentally typed an equal sign when they should have typed a colon. VS Code caught the error, and when the developer hovered over the red squiggle, VS Code explained what the error was and offered to take them to further documentation.

Another useful feature in VS Code (as well as many other code editors) is Git Integration. Instead of using a Terminal window, you can just make a few clicks and easily integrate Git into your workflow!

From VS Code you can:

- Easily see modifications to your code.
- Create a branch.
- Upload your changes directly to GitHub.
- Download changes from other team members to your local system.

IDE Example: Rstudio – IDE

While Visual Studio Code is a more generic IDE where you can use plugins to specialize it, there are also IDEs, such as RStudio, that have specialized features for specific languages right out of the gate.

Researchers conducting statistical analysis tend to use the coding languages of R and Python. RStudio has built-in tools for that very purpose, including data visualization.

Source: https://en.wikipedia.org/wiki/File:RStudio_IDE_screenshot.png

10.6.2 Plain Text Editors for Coding

Most laptop or desktop computers that run standard operating systems (Windows, MacOS, Linux) have multiple pre-installed plain-text editors that can be used for coding. It is beneficial to know how to use at least one, because it makes editing scripts and files a quick process.

PROS

CONS

Lightweight

Many distributed natively with OS

No plugins to help find bugs, errors, etc.

May not have syntax-highlighting

10.6.3 Computational Notebooks

A computational notebook refers to a virtual, interactive computing environment that combines code execution, documentation, and data visualization in a single interface. These notebooks are widely used in data science and coding fields. Popular examples include Jupyter Notebooks and R Notebooks. They allow users to write and run code in a step-by-step manner, providing an efficient platform for data analysis, research, and collaborative coding, with the added benefit of integrating rich text (including equations), images, and charts for clear documentation and communication.

Example: Jupyter Notebook and JupyterLab

Jupyter notebooks are open-source web applications that are widely used for creating computational documents. But before we dive into Jupyter Notebooks, we want to make it clear that Jupyter Notebooks are one of many platforms in the Jupyter ecosystem:

- Jupyter Notebook – contained language shell for interactive programming, displaying output inline with inputs
- JupyterLab – an in-browser user interface showing multiple windows for notebooks, terminals, and code editing
- JupyterHub – middleware for running shared interactive computing environments, including JupyterLab and Jupyter Notebook, on shared computing infrastructure (such as the Cloud)

We will use Jupyter Notebook as an example of a computational notebook and discuss how JupyterLab is related to Jupyter Notebook. The following section on computing platforms will discuss JupyterHub.

This screenshot shows an example of a Jupyter Notebook that integrates rich text (with headers and links), equations, code, and the interactive output from those lines of code, including a plot. This screenshot makes it clear why this is called a computational notebook - it resembles a lab notebook that you may have written out by hand in school.

Project Jupyter | Home

Many programming languages are supported by Jupyter. Fun fact: the name "Jupyter" refers to the three core languages supported by Jupyter: Julia, Python, and R.

JupyterLab is a browser-based interactive development environment that supports Jupyter Notebooks, and is designed in a more flexible environment that allows for many useful features. One of these features is Git integration, as we saw for other IDEs like Visual Studio Code.

Since Jupyter Notebooks allow for integration of code with visualizations and text, they can serve as a tool to carry out research projects and create easily shareable computational documents for education, collaboration, or science communication. With rich text capabilities, such as the use of headers, italics, links, and many more, you can create a readable document that contains runnable code. These are just some of the reasons why JupyterLab and Jupyter Notebooks are widely used across many disciplines, including computational research and data science.

10.6.3.1 For more information on Jupyter products and its community, check out their website.

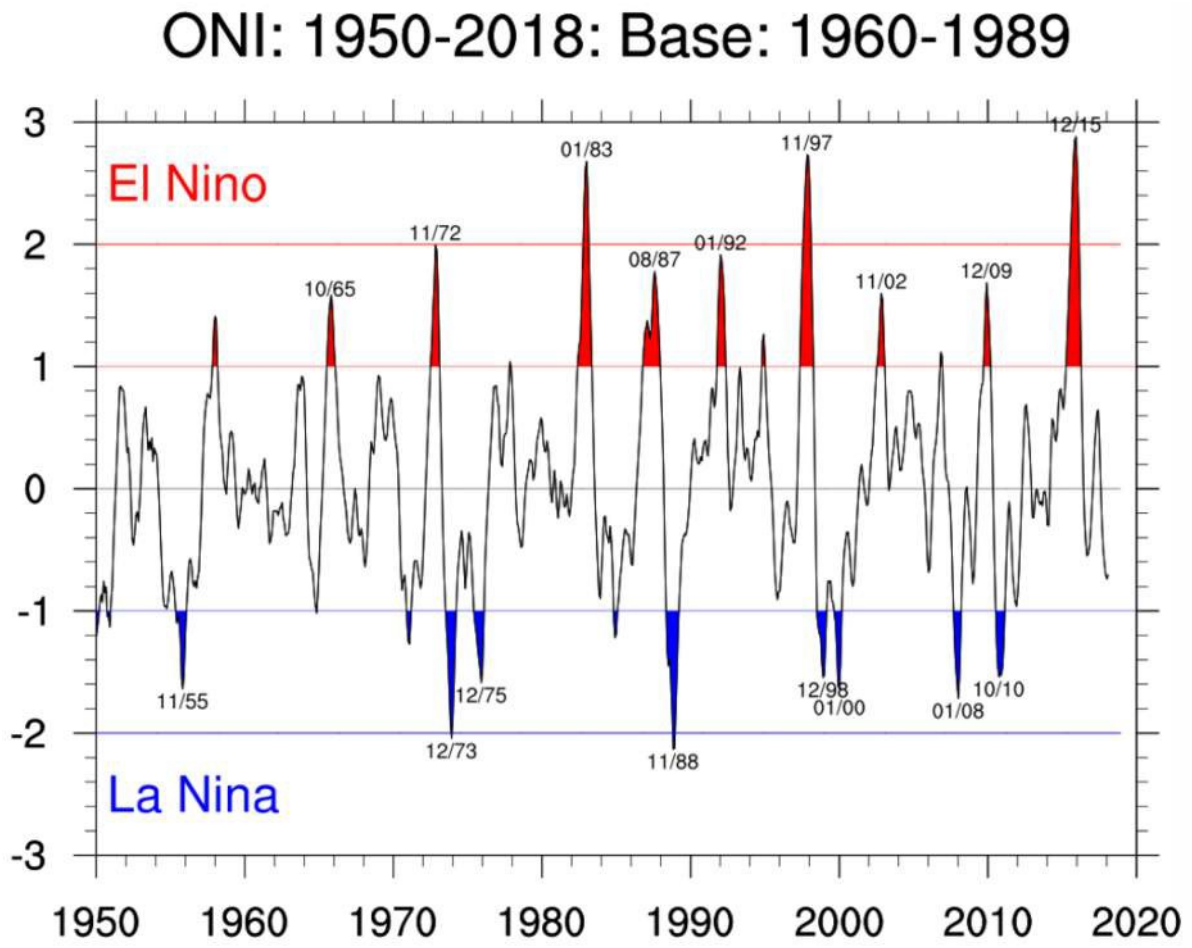
CLICK TO LEARN

If you want to dive in, check out Project Pythia's "Getting Started with Jupyter" lesson, geared toward scientists without assumption of programming background.

10.6.4 Activity 4.1: Run a Jupyter Notebook Yourself from the Browser

Let's use an example from Project Pythia to showcase how computational notebooks can be used in science. Project Pythia is an education Hub for the geoscientific community. They have some great learning resources and example research notebooks that are developed and maintained by the community, and are freely available.

In this activity, you will run pre-written Python code in a Jupyter Notebook from your browser to make plots related to the El-Niño Southern Oscillation (or ENSO). You will use the open-source software package called Xarray to read in sea surface temperature data from a global climate model (the Community Earth System Model - CESM), and create some visualizations of ENSO events across the last 20 or so years. The goal is to recreate the plot below for the last ~20 years. This figure shows the years and magnitude of the El Niño events in red and of the La Niña events in blue.

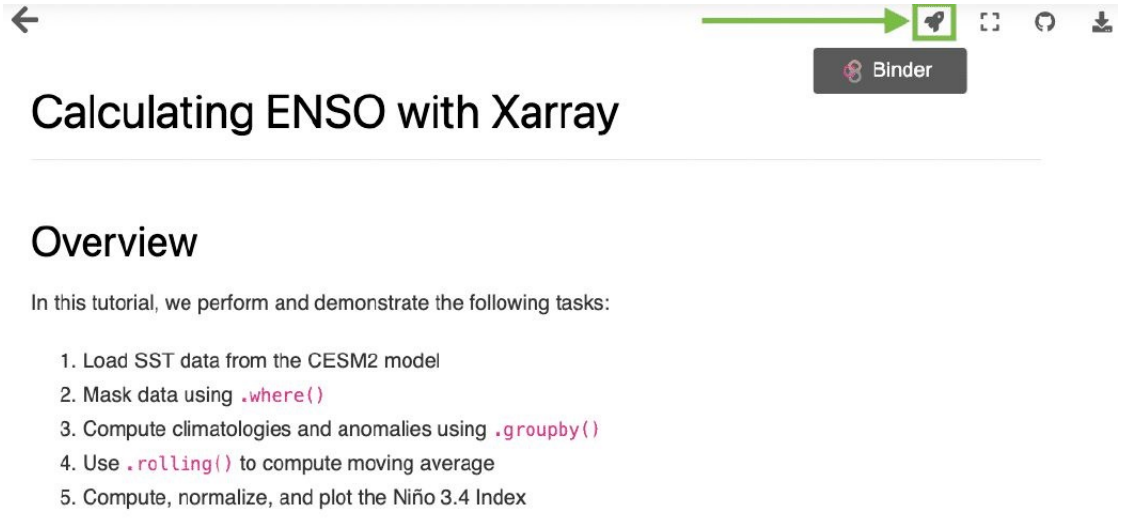


Source: https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni

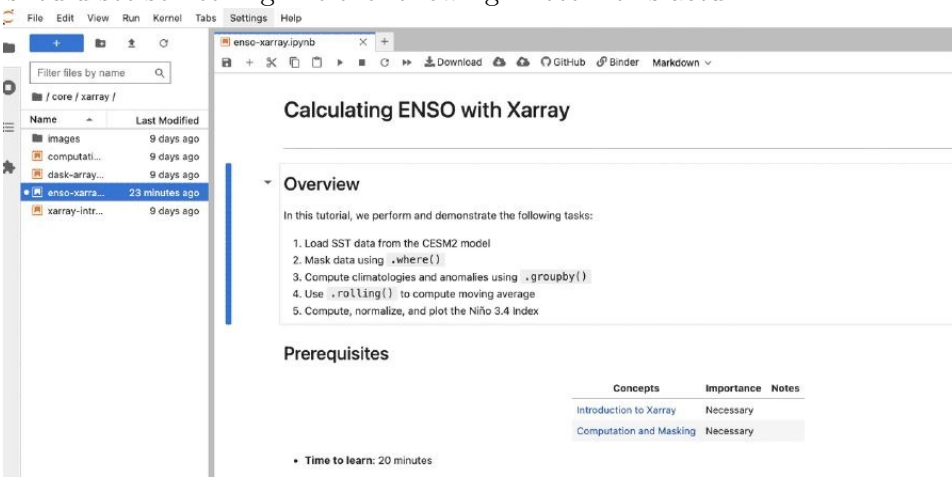
---

Follow These Steps:

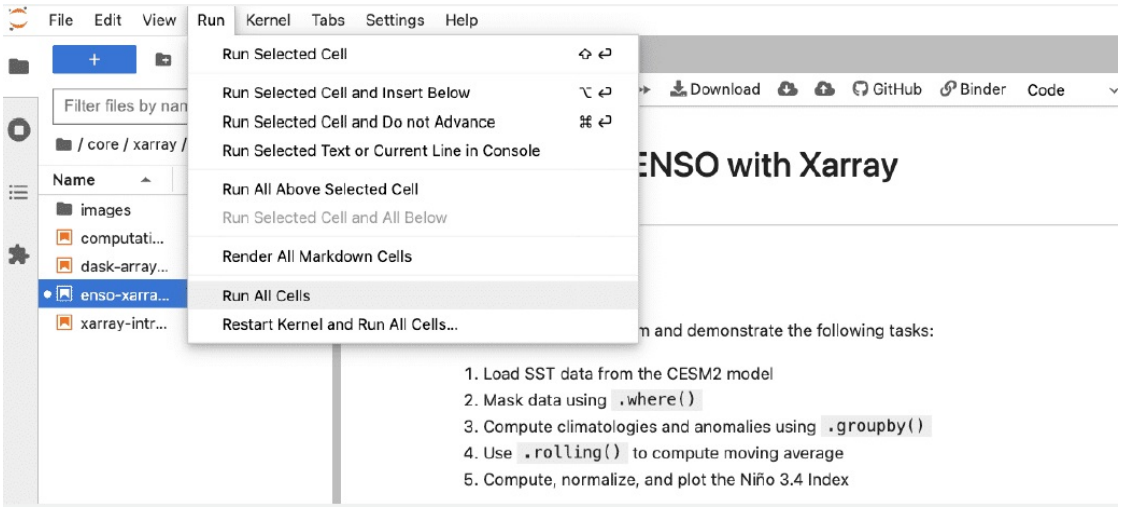
1. Navigate to the "Calculating ENSO with Xarray" lesson
2. In the top right corner, hover your mouse over the rocket icon, and click on "Binder". This will open the lesson as an executable Jupyter Notebook that runs on the Cloud. Note that it may take several minutes for the Notebook to get set up.



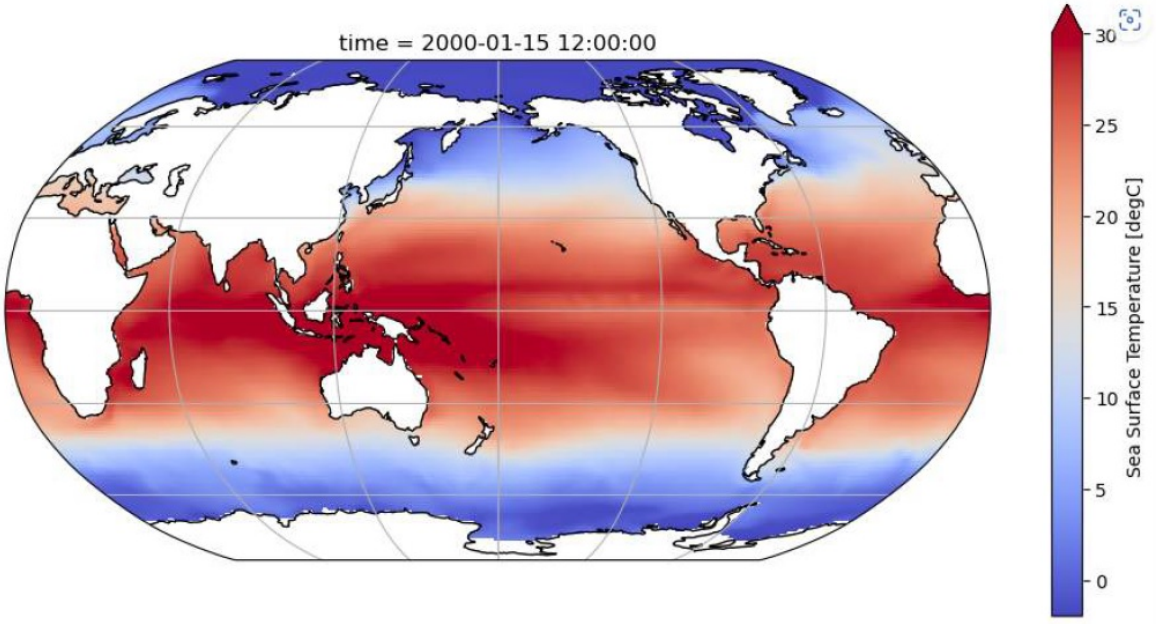
3. After the Notebook loads, you should see something like the following. Note – this actually uses the JupyterLab view!

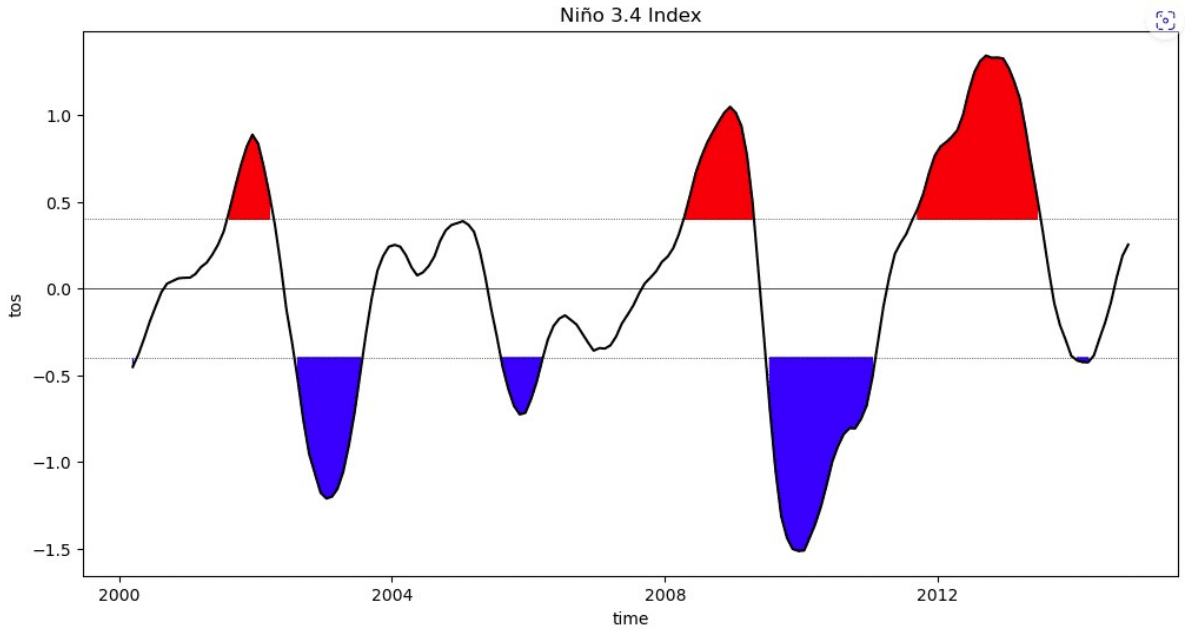


4. You can take a little time to breeze through the text and code in the Notebook, but keep in mind that this lesson assumes a lot of prior knowledge, so it's ok if you don't understand everything. You can still appreciate the nice plots you're about to make!
5. You are now ready to run the notebook yourself! To do that, you can go to the "Run" menu in the upper left of the JupyterLab window and choose "Run All Cells":



6. This should only take a few seconds, and if you scroll down, you can view a couple nice visualizations that you just created: *Use the "<" and ">" buttons to navigate between the images.*





7. Take some time to look through the Notebook a bit more closely. You will see that there is text (including headers, links, and even a table right at the start!), code, and figures integrated together. This is just one example of how scientists use computational notebooks for their research.

You can peruse more of the Project Pythia Python learning resources via their Foundations Book, and you can view more advanced example research workflows in the geosciences that use computational notebooks (which they call "Cookbooks") to see more examples of how notebooks are used in science. If you are interested in the geosciences, you can even contribute your own notebook if you have a notebook you'd like to share!

10.6.5 Computing Platforms

We use the term "computing platform" to refer to the computational machine used to run code. There are many different computing platforms that you can choose, each having their own pros and cons. Here is an overview of three computing options:

10.6.5.1 Personal Computer (e.g. a laptop)

Pros: - Convenient - Can run computations when and where you choose - Can tailor the software environment to be exactly what you need - Don't have to share your computing resources

Cons: - Has limited computational power - Requires downloading data and software

10.6.5.2 High Performance Computing (HPC)

Pros: - High computational power

Cons: - Typically owned and run by a particular institution - may need to be affiliated with that institution to gain access to their HPC - May have to wait significant amounts of time to run your code, since they are typically shared across many people and groups - Need significant funds to build an HPC

10.6.5.3 Cloud Computing

Pros: - Extremely high computational power - Minimal wait times to run code - Typically accessible to anyone with an internet connection - On-demand pricing options - You only have to pay for what you use

Cons: - High cost per computation - Lack of transparency in costs - E.g. it can cost significant amounts to read in data from different Cloud regions, but may not always be clear which region your data and compute are in - May require some extra knowledge in how Cloud computing works

Examples of Cloud providers: - Amazon Web Services (AWS) - Google Cloud - Microsoft Azure

Many data providers, especially of large datasets, are migrating their data to the Cloud to increase accessibility and to make use of the large storage capacity that the Cloud provides. For instance, NASA Earthdata (which houses all NASA Earth science data) is now using AWS to store the majority of its data. Many Cloud providers also have a number of publicly available datasets, including Google Cloud and AWS.

When choosing a computing platform, it is important to consider where your datasets are saved and how big the datasets are. For instance, when working with small datasets, it is often preferable to use a personal computer since data download will take minimal time and large computing resources likely aren't needed. When working with large datasets, however, it is best to minimize the amount of downloading and uploading data that is needed, as this can take significant amounts of time and internet bandwidth. If your large datasets are stored on the Cloud already, it is typically best to use Cloud resources for the computation as well, and likewise for HPC use.

10.7 Additional Tools

10.7.1 Software Repository vs Archive

Software repositories and archives provide centralized locations to store and share software, but there are some important key differences between them that we will discuss in this section.

A software repository is a dynamic and collaborative space where developers work on the latest code, making it the heart of ongoing software development and version control. It houses actively maintained codebases, which encourages collaboration and continuous, often community-driven, improvement.

Conversely, a software archive is static storage where stable and thoroughly tested software releases are kept. Users access these archives to obtain reliable versions of software, ensuring stability and reliability in their applications. Understanding the difference between these two is crucial for effective software development and distribution.

Git/GitHub and Bitbucket are popular choices for software repositories.

Repository

Archive

Is a location for sharing code.

Often use version control systems like Git, Mercurial, and Subversion to track changes

Typically contains the latest development version (sometimes called the “master” or “trunk”) of a software project, which can be actively worked on by developers.

Used for collaborative software development and code sharing among a team or a community of developers.

Important note: A repository is nothing more than a place for hosting code. These days, a version control system and a repository are often one and the same thing. It is important to understand the distinction. However, some websites are purely dropboxes for code executables or zip files of source code.

Repository

Archive

Often used for distribution and long-term preservation of software.

A storage system that contains specific, stable releases or versions of software, compiled binary packages, or source code releases.

Users typically download software from an archive to install and use it on their systems.

Containers

A software container is a standalone, and executable package that includes everything needed to run a piece of software, including the code, runtime, system tools, environment settings, and libraries. Containers are isolated environments that hold the application as well as anything needed to run the application, ensuring consistency and portability across different computing environments. A container is a helpful tool that can provide efficiency, scalability, and ease of deployment. Some examples of widely utilized container tools are Kubernetes, Docker, and Apache Mesos.

10.7.2 Activity 4.2: Match Tools

Match each item to their description:

Integrated development environment (IDE)

Enhanced text editor for code. Assists with identifying syntax and constructs of code

Software archive

Static storage where stable and thoroughly test software releases are kept.

Version control platform

Tool that helps software developers manage and track changes in

10.8 Lesson 4: Summary

In this lesson, you learned: - The usefulness of digital tools that manage, foster collaboration, and house open code. - How version control systems like Git and platforms like GitHub can increase collaboration and management of code. - Some common tools for editing open code, including integrated development environments (IDEs) like Visual Studio Code and Jupyter Notebooks. - The difference between software repositories and archives, and also how software containers can help with the sharing and reproducibility of code.

10.9 Lesson 4: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/03

Which is NOT a benefit to using version control and version control platforms?

- Ability to track changes that have been made

- You cannot go back to make changes
- Ease of code collaboration with others
- Directory clean from clutter - no need for multiple copies of files
- A very targeted backup system for your work

Question

02/03

An interactive computing environment that combines code execution, documentation, and data visualization in a single interface is known as a:

- Version Control Platform
- Software Repository
- Computational Notebook
- Container

Question

03/03

A software repository and a software archive are the same thing.

- True
- False

11 Lesson 5: Tools for Open Results

11.1 Navigation

11.2 Overview

This lesson focuses on the tools available for sharing research products. It begins with a discussion of the tools for management of research projects. Then it introduced the tools for open publications and how to find them. Next, this lesson discusses the tools for open results. Lastly, this lesson discusses the concept of reproducibility. Journals are a tool for sharing your results and these are discussed in more detail in Module 5 - Open Results.

11.3 Learning Objectives

After completing this lesson, you should be able to:

- Describe some of the benefits of preprints and identify resources for open access journals.
- List commonly used tools that increase the reproducibility of a result.
- List applications for project management and reference management.

11.4 Tools for Open Publications

11.4.1 Pre-Prints

Open science tools can be used for writing, as tools to produce content, such as data management plans, presentations, and pre-prints. Pre-prints are early versions of research papers

that are shared publicly before they are published in scientific journals. In some fields, they are shared prior to peer review while in other fields, it may only be after peer review and prior to publication. They are a vital component of open science content creation, as they promote transparency, rapid dissemination of knowledge, and collaboration among researchers.

By sharing pre-prints, scientists can receive feedback from the global research community, refine their work, and rapidly communicate their findings. This accelerates the pace of scientific discovery and ensures that valuable research is accessible to a broader audience, which aligns with the principles of open science.

Pre-prints have gained particular significance during the COVID-19 pandemic, where they played a crucial role in rapidly sharing information about the virus and its effects, emphasizing their importance in advancing science and public health. Fundamentally, pre-prints are important to open science. Consider the following highlights:

1. **Rapid Dissemination:** Pre-prints enable researchers to swiftly share their findings with the scientific community and the public, sometimes within days of completing their research. This swift dissemination is particularly beneficial when dealing with urgent or rapidly evolving topics.
2. **Peer Review:** While pre-prints are not peer-reviewed, they often undergo a form of community review. Researchers and experts can provide feedback and constructive criticism, helping authors improve their work before formal journal publication.
3. **Variety of Fields:** Pre-prints are not limited to any specific scientific discipline. They are used in fields ranging from medicine and biology to physics and social sciences, making them a versatile tool for disseminating research.
4. **Versions and Citations:** Pre-prints can have different versions, and the final peer-reviewed paper may differ. Researchers are encouraged to cite pre-prints when discussing ongoing research, allowing for transparency in the academic discourse.
5. **Free Access:** Pre-prints are typically freely accessible to anyone with an internet connection. This open access promotes equality and inclusivity in science, enabling researchers from various backgrounds and institutions to engage with the latest research.
6. **Not a Replacement for Peer Review:** Although pre-prints are valuable tools for early sharing and collaboration, they are not a substitute for a formal peer-reviewed publication. Researchers and readers should examine pre-prints with the understanding that they have not undergone the rigorous peer review process that journals provide.

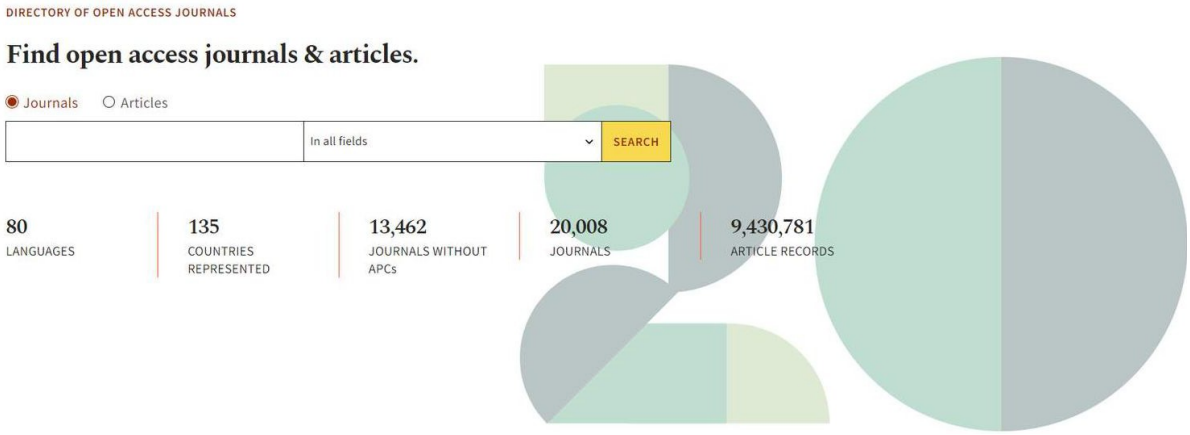
Pre-prints are typically hosted on dedicated pre-print servers for different scientific fields. Examples include: arXiv (physics, mathematics), bioRxiv (biology), medRxiv (medicine), and many others. These platforms help organize and facilitate pre-print sharing. The OSF provides a services for searching over multiple preprint servers.

Remember, pre-prints play a significant role in open science by promoting rapid, transparent sharing of research findings across various scientific domains. They offer a valuable platform for researchers to disseminate their work and gather feedback, ultimately advancing scientific knowledge.

11.4.2 Discover an Open Access Journal to Share Your Results

A common way to share a paper is to pick a journal that is already fully open access and adopt their license. One way to discover open journals is by using the Directory of Open Access Journals (DOAJ).

To identify the best open-access journal, you can use the Directory of Open Access Journals (DOAJ) which provides a searchable index of all known open-access journals and articles. The DOAJ and its synergetic webpage, Sherpa Romeo, serve as useful tools in the early stages of research planning to help a researcher determine what journals to consider when the time comes to publish their results.



11.4.3 Activity 5.1: Identify an Open-Access Journal

To become more familiar with the DOAJ, visit https://doaj.org/ and search for *The Astronomical Journal* published by the American Astronomical Society. Once you select the journal, you can see costs to publish, details about licensing, author retention rights, time to publication, and other details.

Once you have found the journal, answer the following questions:

1. When did it begin publishing as open access?
2. What license is used for the publications?
3. What rights do the authors retain in their publications?

Note: If journals did not have any open access, the journal will not have appeared in the search results. Also, because DOAJ has strict criteria for being listed in its directory, it is not likely you will find predatory publishers listed here, either.

11.5 Tools for Reproducibility

In this lesson, we take a deep dive into a few available tools for (computational) reproducibility.

11.5.1 What is Reproducibility?

The National Academies Report 2019 defined reproducibility as:

- Reproducibility means obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis.
- Replicability means obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

The pursuit of reproducibility aims to ensure researchers reach the same result when using the same steps, as well as to enable researchers to copy an environment and build upon a result by editing the environment in order to apply it to a similar problem. This additional feature gives others the ability to directly build upon previous work and get more science out of the same amount of funding.

Tools to support reproducibility in research outputs:

- Jupyter Notebooks - A web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience.
- Jupyter Books - Build beautiful, publication-quality books and documents from computational content.
- R Markdown - Produces documents that are fully reproducible. Use a productive notebook interface to weave together narrative text and code to produce elegantly formatted output.
- Binder - Create custom computing environments that can be shared and used by many remote users.
- Quarto - Combine Jupyter notebooks with flexible options to produce production quality output in a wide variety of formats.

Note: As you might have noticed, a lot of open science tools require intermediate to advanced skills in data and information literacy and coding, especially if handling coding-intensive research projects. One of the best ways to learn these skills is through engaging with the respective communities, which often provide training and mentoring.

11.6 Additional Tools for Open Results

11.6.1 Tools for Open Project Management

Advancements over the past few decades to tools that manage research projects and laboratories have helped to meet the ever-increasing demand for speed, innovation, and transparency in science. Such tools are developed to support collaboration, ensure data integrity, automate processes, create workflows and increase productivity.

Research groups can now use project management tools for highly specialized efforts. They use existing platforms or develop their own software to share materials within the group and manage projects or tasks.

Platforms and tools, which are finely tuned to meet researchers' needs (and frustrations), are available. They are often founded by scientists, for scientists. To explore a few examples, let's turn to experimental science.

A commonly used term and research output is protocol. Protocol can be defined as "A predefined written procedural method in the design and implementation of experiments. Protocols are written whenever it is desirable to standardize a laboratory method to ensure successful replication of results by others in the same laboratory or by other laboratories." according to the University of Delaware (USA) Research Guide for Biological Sciences.

In a broader sense, protocol comprises documented computational workflows, operational procedures with step-by-step instructions, or even safety checklists.

Protocols.io is an online and secure platform for scientists affiliated with academia, industry and non- profit organizations, and agencies. It allows users to create, manage, exchange, improve, and share research methods and protocols across different disciplines. This resource can improve collaboration and recordkeeping, leading to an increase in team productivity and facilitating teaching, especially in the life sciences. In its free version, protocols.io supports publicly shared protocols, while paid plans enable private sharing, e.g. for industry.

Some of the tools are specifically designed for open science with an open-by-design concept from ideation on. These tools aim to support the research lifecycle at all stages and allow for integration with other open science tools.

As an example, the Open Science Framework (OSF), developed by Center for Open Science, is a free and open source project management tool. The OSF supports researchers throughout their entire project lifecycle through open, centralized workflows. It captures different aspects and products of the research lifecycle, including developing a research idea, designing a study, storing and analyzing collected data, and writing and publishing reports or papers.

The OSF is designed to be a collaborative platform where users can share research objects from several phases of a project. It supports a broad and diverse audience, including researchers

that might not have been able to access certain resources due to historic socioeconomic disadvantages. The OSF also contains other tools in its own platform.

"While there are many features built into the OSF, the platform also allows thirdparty add-ons or integrations that strengthen the functionality and collaborative nature of the OSF. These add-ons fall into two categories: citation management integrations and storage integrations. Mendeley and Zotero can be integrated to support citation management, while Amazon S3, Box, Dataverse, Dropbox, figshare, GitHub, and oneCloud can be integrated to support storage. The OSF provides unlimited storage for projects, but individual files are limited to 5 gigabytes (GB) each."

Center for Open Science

11.6.2 Best Practices for a Project Registry

It is common for different types of outputs to be preserved in different places to optimize discovery and reuse. An up-to-date Project Registry provides a quick overview of all the outputs. Best practices for managing a Project Registry include:

- Create and update a Project Registry in conjunction with preserving outputs (as described above) in the form of a spreadsheet or other type of list. This can be one registry for the entire project that is updated, or a new registry for each milestone.
- Include in each registry entry a description of the object, preferred citation, and the persistent identifier (e.g., DOI), and any other useful information supporting the project. For outputs that do not have a persistent identifier, provide a URL and description.
- Preserve the Project Registry as a project component. Many funders require in their yearly reports a list of both peer-reviewed publications and all project outputs. The Project Registry can be provided to the funder during the reporting process, or used as a tracking tool to assist with completing the report.

11.6.3 Managing Citations Using Reference Management Software

Keeping track of every paper you reference, every dataset you use, and every software library you build off of is critical. A single paper might cite dozens of references, and each new thing you produce only adds to that list. Reference Management Software can be employed to help you manage these references and automatically create a list of citations in whatever format you need (BibTeX, Word, Google docs, etc.).

While you are writing up results, keeping track of references and creating a correctly formatted bibliography can be overwhelming. A management software can keep track of references and can be shared with colleagues who are also working in the document.

Some of the common capabilities of reference management software are:

- Keep database of article metadata
- Import article metadata from PDFs
- Track datasets and software versions and DOIs
- Create formatted references and bibliography for many different journal styles

Examples of reference management software include:

- Mendeley
- EndNote
- Zotero

11.6.4 Open Highlight: Zotero

Zotero helps manage software, data, and publication metadata and citations through a drag-and-drop interface. Researchers can use the tool to automatically generate citation files (for example, in BibTeX format).

- Open Source
- Drag and Drop PDFs to import metadata
- Word + Browser plugins
- Export citations to BibTeX

11.7 Lesson 5: Summary

In this lesson you learned:

- Benefits of preprints and resources for open access journals.
- Tools for reproducibility and replication of your studies.
- Additional tools that are available to help manage open results including project management and reference management.

11.8 Lesson 5: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/03

Read the statement below and decide whether it's true or false:

Reproducibility means obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis.

- True
- False

Question

02/03

Which of the following steps would you take to manage a project registry for outputs? Select all that apply.

- Create and update a Project Registry in conjunction with preserving outputs in the form of a spreadsheet, or other type of list.
- Include in each registry entry a description of the object, preferred citation, and the persistent identifier (e.g., DOI), and any other useful information supporting the project.
- Preserve the Project Registry as a project component.
- Log all outputs in a paper notebook

Question

03/03

Which is NOT a capability of Reference Management Software?

- Keep database of article metadata
- Import article metadata from PDFs
- Track datasets and software versions and DOIs
- Create summaries of research articles

11.9 Open Tools and Resources Summary

Throughout this module, you learned about some of the concepts and tools that support the discovery and use of open research, that can be used to make data and software, and that can be used to share your results. These included:

- The foundational elements of open science, which includes research products such as data, code, and results.
- Resources used to discover and assess research products for reuse, including repositories, search portals, publications, documentation, metadata, and licensing.
- Making and sharing data that employs the FAIR principles by incorporating a data management plan, using persistent identifiers and citations, and utilizing the appropriate data formats and tools for making data and sharing results.

- The use of the tools needed for development of software including source code, kernels, programming languages, third-party software and version control.
- The tools and documentation types used for publishing and curating open software.
- Resources for sharing research products including preprints, open access publications, reference management systems, and resources to support reproducibility.

Part III

OS101 Module 3: Open Data

Welcome to Open Science 101: Open Data

About This Module

This module focuses on the practice and application of open science for data. It provides a 'how to' process for finding and assessing open data for use, for making open data and for sharing open data. The step-by-step flows are easy to follow and can be used as checklists after you complete the module. Some of the key topics discussed include: data management plans, the process for assessing data for reuse, creating a plan for making data including choosing open formats and adding documentation, and the considerations for sharing data and making your data citable.

Module Learning Objectives

After completing this module you should be able to:

- Describe the meaning and purpose of open data, its benefits, and how FAIR principles are used.
- Recall methods to assess the reusability of data based on its documentation, and cite the data as instructed.
- Implement an open data management plan, select open data formats, add the needed documentation, including metadata, README files and version control, to make the data reusable and findable.
- Evaluate whether your data should and can be shared.
- Recall practices to make data more accessible, including the registration of an affiliated DOI and the inclusion of citation instructions in documentation.

Key Terms

These key terms are important topics for this module. Select the term to see the description.

Copyright – A type of intellectual property that protects original works of authorship as soon as an author fixes the work in a tangible form of expression. Many different types of works are covered by copyright law including data products and software. (As well as books, poems, paintings, photographs, illustrations, musical compositions, and many more.)

Data – Any type of information, recordable, or observable facts. Data are now most commonly stored electronically.

Data License – Data licenses give any data creator a way to grant the public permission to use their products under copyright law. Similarly, data licenses give data users clear guidelines regarding how they can reuse the material.

CC-BY and CC0 License – CC-BY and CC0 are Creative Commons data licenses. CC-BY allows reusers to distribute, remix, adapt, and build upon the material in any medium or format so long as attribution is given to the creator. The license allows for commercial use. CC0 allows creators to give up their copyright and put their works into the worldwide public domain. CC0 allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, with no conditions.

Data Management Plan – A data management plan (DMP) describes the what, where, when, and who for data that will be created during a research project. Common components of data management plans include a description of the type, volume, and format of the data; where and when the data will be made available, and who will make the data available. The plan can also describe data variables, sources, accuracy, and precision if that information is available.

Metadata – Data that describes data. It can be global – describing the overall contents of a single file or collection of files – or local – describing an individual variable within the file. Typically, global metadata offers information about who created the file, information about the data set, what satellite/instrument/lab/etc. created the set, the DOI, and file format information, among other metadata fields. Local metadata about variables contains information such as the full/long name of the variable, any scaling factors or uncertainty information, and measurement units.

Machine-Readable Persistent Identifiers (PID) – A unique string that identifies an object, such as a dataset. Though the online location of the object may change, the PID will not, and will also lead back to the data, ensuring that citations referencing the PID will always be valid.

Findable (data) – Data that is readily discoverable to both humans and machines. It should include a unique persistent identifier and rich metadata describing the data and context and be registered in an index that is searchable.

Accessible (data) – Data that can be accessed over standard communication protocols, with metadata that can be accessed even if the data itself is no longer available.

Interoperable (data) – Data that uses controlled ontologies and vocabularies so that it can be used and/or combined with other relevant data sets in different applications.

Reusable (data) – Data that has a clear license, detailed provenance, adequate description/definition and meets community/domain standards, and can be replicated or combined with other data.

Dataflow – The data workflow that includes how data are used, made, and shared. Different actors will have different (or multiple) roles in this workflow.

Navigation

Lesson 1: Introduction to Open Data

Lesson 2: Using Open Data

Lesson 3: Making Open Data

Lesson 4: Sharing Open Data

Lesson 5: From Theory to Practice

12 Lesson 1: Introduction to Open Data

12.1 Navigation

12.2 Overview

This lesson defines open data, its benefits, and the practices that enable data to be open. In addition, the lesson takes a closer look at how FAIR applies to open data as well as at the criticall role of metadata. It wraps up with a brief discussion on how to plan for open data in the scientific workflow and tasks guided by the use, make, share framework.

12.3 Learning Objectives

After completing this lesson, you should be able to:

- Define what open data is and how the FAIR and CARE principles are used to guide open data practices
- List the benefits of open data
- Explain how the use, make, share framework can be used to modify the scientific plan for open data

12.4 Introduction

Data drives science forward. Data are stored electronically to enable further analysis and research. Digital technologies integrated into every aspect of modern scientific research has led to the production of large amounts of data.

Open data is an essential pillar of open science. In many ways, open data are a natural expansion of open science beyond scholarly publications to include digital research outputs. It has since become an integral part of the open science movement as open data allows anyone to see, use, and verify published results. Open data makes science more accessible, inclusive, and reproducible. In order to support this, data needs to be made available in formats that others can use, include metadata that describes the data, and provided with helpful documentation. When made available, open data enables new discoveries and unforeseen uses.

12.4.1 Example: How Will Humans Live on the Moon or Travel to Mars When the Space Environment Threatens Human Health in Multiple Ways?

Bone loss, vertigo, anemia, muscle atrophy, increased risk for cancer - these are just some of the human side effects of space travel. To study these human health risks of space travel, scientists around the world use NASA's open-source GeneLab platform. GeneLab aggregates large volumes of space biology data on human and model organism samples exposed to spaceflight conditions. Their digital and physical repositories include cell info as well as DNA, RNA, and proteins. As an open-source platform, GeneLab data are publicly accessible at no cost.

Example: Using astronaut biological data from GeneLab, scientists recently found what may be the culprit behind many of the side effects from travel to space: mitochondrial stress.

Watch Video

Mitochondria are components within our cells that affect respiratory and energy function. This discovery could be crucial to overcoming human health- related problems in space. Understanding the source of this issue could help scientists develop countermeasures and therapies to keep people healthy in space for longer periods of time.

12.5 Definition and Considerations of Open Data

12.5.1 What is Data?

The Turing Way Community. This illustration is created by Scriberia with The Turing Way community, used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807

Data are any type of information that is collected, observed, or created in the context of research. Today, data are increasingly stored electronically in a digital format.

Data includes:

Primary (raw) data – Primary data refers to data that are directly collected or created by researchers. Research questions guide the collection of the data. Typically, a researcher will formulate a question, develop a methodology and start collecting the data. Some examples of primary data include:

- Responses to interviews, questionnaires, and surveys.
- Data acquired from recorded measurements, including remote sensing data.
- Data acquired from physical samples and specimens form the base of many studies.
- Data generated from models and simulations.

Secondary & Processed data – Secondary data typically refers to data that is used by someone different than who collected or generated the data. Often, this may include data that has been processed from its raw state to be more readily usable by others.

Published data – Published data are the data shared to address a particular scientific study and/or for general use. While published data can overlap with primary and secondary data types, we have "published data" as its own category to emphasize that such datasets are ideally well-documented and easy to use.

Metadata – Metadata are a special type of data that describe other data or objects (e.g. samples). They are often used to provide a standard set of information about a dataset to enable easy use and interpretation of the data.

The term open data are defined in the open data handbook from the Open Knowledge Foundation:

"Open data are data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and share alike."

Open Data Handbook from the Open Knowledge Foundation

When talking about data in the context of this module, we focus on the data that you are preparing to share, such as data affiliated with a scientific publication, regardless of what type that is. While you could share (and many do) laboratory notebooks, preliminary analyses, intermediate data products, drafts of scientific papers, plans for future research and similar items, these aren't usually required by funding agencies or institutions and thus won't be in focus for this module.

To quote from a published paper about data reuse, researchers are mostly looking for data which is "comprehensive, easy to obtain, easy to manipulate, and believable." For these criteria to be fulfilled, the data should:

- Be sufficiently described with appropriate metadata, which greatly affects open data reusability. There is no one size fits all for metadata as its collection is guided by your data.
- Have the appropriate license, copyright, and citation information.
- Have appropriate access information.
- Be findable in an accredited or trustworthy resource.
- Be accompanied with history of changes and versioning.
- Include details of all processing steps.

Not all data may be shared or shared with all this information. There are different reasons why it might not be possible. However, the more information shared about data helps increase the reliability and reusability of the information.

12.6 Benefits of Open Data

Data underpins almost all of science. Openly sharing data with others enables reproducibility, transparency, validation, reuse, and collaborations. Data plays a significant role in our day-to-day lives. Open data, in particular, plays a key role. Open data are only common in our society and you have likely already benefited from this form in some way. The impacts of open data include facilitating:

Greater Good – Data plays a significant role in our day-to-day lives. Open data, in particular, has played a key role. If you pause and think about it, you may realize that open data are not only common in our society, but you might have benefited from and used open data yourself.

Each country or territory often provides open access to a variety of socioeconomic information about the population, community, and business in its jurisdiction. These data are often called census survey data which may include the aggregated statistics of gender, race, ethnicity, education, income, and health data of a community. These data are often used to understand the composition of a local neighborhood and are critical to inform decisions on resource allocation to ensure the quality of life for the community.

12.6.0.1 Example: Open Data Helps Provide Life-Saving Information in the Face of Climate Change

The changing climate poses a significant risk to our daily lives and has been responsible for intensifying drought, increasing flooding and devastating fire incidents worldwide. Open data are therefore critical in providing life-saving information to adapt to the changing climate and help assess the climate risks where we live. Government agencies have been providing public access to long-term weather and climate information for decades (e.g., National Oceanic Atmospheric Administration in the U.S., UK Met Office, European Centre for Medium-Range Weather Forecasts). A more recent initiative stems from organizations developing value-added open-data products to advise society on the risk of a changing climate. One recent example is the flood and fire risk in the United States developed by a non-profit organization First Street Foundation.

Policy Change

12.6.0.2 Example: Predicting Climate Change Effects in Arctic Communities

Open data can lead to policy change that directly impacts the lives of communities, such as those destined to suffer first from the slow changes to the Arctic. A study in Nature employed OpenStreetMap data to help produce maps of projected environmental changes in the Arctic. These maps helped emphasize the need for adaptation-based policies at community and regional levels to avoid stagnation of change in the light of a sudden and dramatically worsening situation fueled by climate change.

Global Emergency Response

12.6.0.3 Example: COVID-19

The COVID-19 pandemic demonstrated to the world, in real-time, how the collective movement of researchers sharing their data (such as sharing of coronavirus genome data) can lead to an unprecedented number of discoveries in a relatively short amount of time. This directly impacted radical vaccine development efforts and the timely control of the COVID-19 infection. These insights will continue to pay off, with this research spurring future developments.

Data sharing has many benefits and can aid access to knowledge. However, it is important to consider where the data has come from, who should have a say in its interpretation and use, and how the data can be shared responsibly.

Citizen Science

12.6.0.4 Example: Water Quality Testing in Beirut

A citizen scientist is a citizen or amateur scientist who collaborates with professional researchers to help gather or interpret data on a broader spatial and temporal scale than the researchers might be able to achieve on their own. This outsourcing of responsibility helps members of the public engage in scientific pursuits that ultimately benefit them and allow research to be conducted on a grander scale than that might be possible with only professional researchers. Citizen science is gaining popularity and recognition as a valuable contribution to scientific advancements.

For example, volunteer citizen scientists in Beirut were recruited from 50 villages to help test water quality [cite: chapter 5 of Contextualizing Openness: Situating Open Science]. These volunteers were trained to be able to conduct the tests and in turn, not only was the data collected to inform the scientific advancements, the citizen scientists had the opportunity to learn to better manage their water resources and were able to improve conditions, creating a mutually beneficial interaction.

Open Data and Equitable Sharing of Knowledge

Free distribution of knowledge increases participation in science. Open data are central to fostering science that is inclusive and diverse, with direct and relevant benefits to impacted individuals and communities. This integration with communities is particularly important in the mission towards the equitable sharing of knowledge.

In a research ecosystem where knowledge is a commodity, with the main currency in the form of published papers and hoarded datasets, exclusion from research can limit scientific progress and negatively impact community outcomes. Those excluded from traditional science resources are often from low and lower middle income countries. Opening our data in an inclusive and easily reusable way is one step toward purposeful inclusion of underrepresented groups in science.

12.6.0.5 Example: Recognition and Compensation for the Work of African Ebola Researchers

During the West African Ebola outbreak from 2014-2016, West African researchers actively worked to collect blood sample data to better understand the Ebola virus and to help put a stop to the rapid spread of the virus. However, most of the blood samples were sent overseas to the US and Europe, where researchers used those data samples to author papers about Ebola. According to the paper "Science under fire: Ebola researchers fight to test drugs and vaccines in a war zone", "This frustrated researchers in the countries ravaged by the virus, who had hoped that studying aspects of the epidemic would strengthen their ability to respond to future infectious- disease outbreaks."

By fostering a global research culture of transparency and validation, where the work of underrepresented groups is celebrated and compensated, we will create a sustainable model that ensures under-represented communities (such as women, under-represented communities, indigenous scholars, non- Anglophone scholars) a voice in how the global and nuanced narrative of science is developed.

Open data that are purposefully inclusive and open to scrutiny, benefit scientific innovation by allowing for a more diverse and robust scientific process that draws from multiple perspectives. This openness also allows for the early identification of mistaken insights as well as early intervention for unforeseen harms to impacted communities.

Open data allows non-traditional researchers to contribute to scientific development and bring their unique insights to the table. With these benefits in mind, we should always bear in mind that Open Data requires careful consideration of its potential downsides that results from failure to provide due credit and consultation with potentially vulnerable and/or marginalized communities. The next lesson "Using Open Data" discusses important considerations for the responsible management, collection, and use of open data by all stakeholders.

12.6.1 Benefits to You

Open data also benefits your research and career. For starters, you are your own future collaborator!

Doing open science not only lets other people understand and reproduce your results, but lets you do so as well! Implementing open science principles like good documentation and version control helps you, potential collaborators, and everyone else to understand your results. In 2 hours, 2 weeks, or 2 years, you will still be able to understand what you did.

Specific benefits of opening data for you as an individual:

- You will never lose access to your previous work, no matter what institute you are affiliated with. Many researchers move around institutions and organizations and by having your data publicly accessible in repositories, you will always have access to them.
- Your data can be cited and you will get credit.
- Publications that include links to data are cited more, according to a 2020 study.

Implementing best practices for open science can strengthen your funding proposals. Funding agencies are realizing that openly sharing research provides more return on their investment. Well-documented research products also demonstrate the quality of your work, which helps with public communication and can also attract quality collaborators. Everybody prefers to work with people who are reliable and do a good job.

12.6.2 Activity 1.1 Open Data Review

Take a moment to reflect on what data sharing means to you.

Image source: CC-by openaire

The word cloud showcases the variety of meanings and interpretations that people have about open data. How many terms in the word cloud do you recognize? Are any of them new to you?

12.7 Challenges of Open Data

While open data has many benefits, there can also be challenges to its creation and use. Throughout this Module, we discuss many of these challenges and possible solutions. In this section, we discuss a few of the most common concerns along with actions to mitigate them.

Example: Are There Any Harms to Open Data?

Open data has been demonstrated to further marginalize or exploit small- scale and community driven initiatives, such as in the case of African researchers neither receiving due credit nor compensation for their genome sequencing during the COVID-19 pandemic. This is further explored in the next section as we introduce ways of mitigating harms that could happen via unthoughtful and irresponsible sharing of data.

12.7.1 Restrictions on Sharing Data

Some data should only be shared very carefully or not at all. Reasons not to share can include:

- Data includes a country's military secrets or violations of national interests.
- Data includes private medical information or an individual's personally identifiable data.
- Indigenous/cultural/conservation concerns.
- Data includes intellectual property.

It is important to be familiar with the policies around sharing of your data and policies from your funding agency, institution, or laws around data protection. These are further discussed in later modules.

12.7.2 Common Fears Around Sharing Open Data

12.7.2.1 NOTE: We will discuss many of the concepts mentioned in the discussion/mitigation column later in this module.

Fear: Scooping: What if someone re-uses my data to publish a result I was working on?	Yes, this can happen. But, in many fields, if it is clear that someone is actively working on a problem, the decision by another to scoop may have a short term gain but long-term loss. In science, reputations are very important and being collaborative generally leads to increased career successes. If you are sharing your data, ensure it has a digital object identifier (DOI). This does not prevent someone from using your data without attribution, but it helps make it easy for others to cite your data. There is a nice article about this here.
Fear: Misinterpretation or Misuse	Provide sufficient contextual information (documentation) to allow others to understand your data fully to reduce this risk.
Fear: My data will be used but not cited	While it is not common for researchers to cite data, science ethics dictates that you should be cited if your work is used. And remember to cite others' data, so you're not adding to the problem!
Fear: Data are too sensitive to share	Use controlled access to help maintain sensitivity and security.
Fear: My data won't be useful to anyone else	You never know how materials might be used! Sailors in the 1800s collected temperature data that is an important part of our ocean climate record today!

These are all valid concerns when sharing data openly, but as indicated by the global move towards open science, the overall benefits outweigh the concerns.

Ultimately, you are free to deploy the open data principles and resources in your research to maximize its impact and meet the expectations of your sponsors and community while managing costs.

12.8 Applying FAIR Principles

Image by Patrick Hochstenbach, CC0 1.0; image illustrates the each FAIR principle

---

12.8.1 FAIR: Findable, Accessible, Interoperable, Reusable

The vast majority of data today is shared online. FAIR principles help researchers make better use of, and engage with a broader audience with, their scientific data than outdated techniques would allow. FAIR data are more valuable for science because they are easier to use. Data can be FAIR regardless of whether it is openly shared or not. If data are openly shared, being FAIR helps with reuse and expands the scientific impact of the data.

FAIR principles don't encompass comprehensive implementation instructions for every type of data, but offer general insights to improve shareability and reusability. Sometimes it takes a group effort and/or a long production process to make data and results FAIR. The process starts in the planning stage of a research project. A well-coordinated open science and data management plan is often needed for full compliance with FAIR, depending on the size and type of project the data are used for.

Up-to-date information about FAIR Principles can be found at the GO FAIR Initiative website

CLICK TO LEARN

Let's review how to make data FAIR for your community.

Select each tab to find out more information.

FINDABLE

ACCESSIBLE

INTER-OPERABLE

REUSABLE

To ensure that data are findable by those in your community:

Deposit data in repositories to preserve the data over time.

Assign your dataset a persistent identifier (PID), such as a digital object identifier (DOI).

Add rich, self-describing metadata in your data files and register the metadata in a metadata catalog that will enable your data to be properly curated.

```
<ul>
            <li>Note that some images or binary files cannot be readily indexed or searched a
            </ul>
        <li>Automate the sharing of your metadata with targeted communities, if applicable.</
        </ul>
    </td>
</tr>
```

FINDABLE

ACCESSIBLE

INTER-OPERABLE

REUSABLE

To ensure that data can be accessed by those in your community:

Archive in a data repository/data center with standardized access protocols.

Repository access protocols should be well-defined and ideally should support machine-to-machine access.

Provide information on how users can access your data, ideally in an automated, machine-based fashion.

If the full content cannot be made openly available for any reason (data sensitivity, infrequent data access, file storage issues), the metadata can still be made openly available so that users can find out who they need to contact to request the data (if possible).

FINDABLE

ACCESSIBLE

INTER-OPERABLE

REUSABLE

To ensure that data are interoperable for those in your community:

Report the data in community standard format.

Use existing standardized metadata if available to minimize “lost in translation” issues and support machine-readability.

The use of controlled terminologies, vocabularies, and ontologies is necessary to support interoperability, but may not yet be available in all research fields.

FINDABLE

ACCESSIBLE

INTER-OPERABLE

REUSABLE

To ensure that data are reusable by those in your community:

Ensure that metadata accurately describes the data and its variables as well as any particularities or limitations.

Specify clear usage licenses for your data.

Provide accurate information on provenance in your metadata.

Add enough metadata information so that your data can be properly cited when it is used.

12.8.2 Metadata's Central Role in Applying FAIR

Metadata are important for search engines to find data and for people to be able to easily compare what is returned.

- Metadata are essential to the implementation of FAIR Principles and enable the data to be used by machines in an automated fashion.
- The richer and more self-describing metadata are, the better they will be handled by anyone who is interested in your data.

12.8.3 Licensing Data

A license is a legal document that tells users how they can use a particular dataset. If you don't license your dataset, others can't/shouldn't re-use it - even if you want them to! It is imperative to understand the licensing conditions of a dataset before data reuse. Without a good understanding of what a license allows, data users may run into copyright infringement or other intellectual property issues.

To ensure open reuse of your data, you can use an open license. An open license has language that describes the user's ability to access, reuse and redistribute the dataset. There are many types of data licenses that are open to varying degrees, and these will be discussed further in the lesson "Making Open Data".

12.9 Planning for Openness: Using the Use, Make, Share Framework for Open Data

12.9.1 Open Science and Data Management Plans

Most scientific funding agencies and organizations ask for a plan for sharing your research when you propose a project. One example of an open science plan is the Open Science and Data Management Plan (OSDMP) for NASA's Science Mission Directorate (SMD) that describes how the scientific information that will be produced from scientific activities will be managed and made openly available. The OSDMP includes sections on data management, software management, and publication sharing; the latter two will be discussed in future modules. If your study has other types of outputs, such as physical samples, hardware, or anything else, you should include those in the plan too.

A best practice when beginning your open data journey is to create a Data Management Plan, or DMP which goes within the OSDMP. This describes how you will manage, preserve, and release data, during and after a research project. Common elements of DMPs relevant to open data include a description of the following:

What?	Data types, volume, formats, and (where relevant) standards.
When?	The schedule for data archiving and sharing.
Where?	The intended repositories for archived data.
How?	How the plan enables long- term preservation of the data.
Who?	Roles and responsibilities of the team members in implementing the DMP.

Investigate if your home institution or funding source has guidance, standards, or templates for DMPs. Other organizations have DMP guidelines and examples as well:

- USGS
- NOAA
- NSF

More details about how to create these plans will be provided in the lesson "From Theory to Practice".

12.9.2 Scientific Workflow

There are a variety of scientific workflow models that explain open science. Data plays a central role in the scientific workflow, where users can propose to create new data, collect and package their data during their project, then archive it for long term storage/use/reuse.

For this curriculum, we use the workflow model from Opensciency. It is used to illustrate that regardless of the workflow model you use, the adoption of open data is performed throughout the entire workflow and production of associated deliverables.

If your project is already in progress, it is a good idea to update future data releases to adhere to open data principles as much as possible. For new projects, your proposals should include creating open data from the start of your project.

In this curriculum, content is organized by how you might use it, make it, and share it. Part of doing open science is building on others' materials (using), creating materials yourself (making), and sharing those so others can use those results (sharing). The lessons are all organized around these steps in the scientific workflow.

The "Use, Make, Share" framework categorizes the tasks commonly used in the practice of open science.

12.9.3 Roles in Use, Make, Share

Individuals interacting with data at various points in the scientific workflow can take on different roles. It is possible that these roles can overlap depending on project requirements, the size of your team, and even funding. All should be using open data principles to perform their tasks. Generally, roles include:

Select each tab to find out more information.

DATA USERS

DATA MAKERS (DATA PROVIDERS)

DATA SHARERS (DATA PUBLISHERS)

Data users primarily discover, assess, and utilize data in research projects.

DATA USERS

DATA MAKERS (DATA PROVIDERS)

DATA SHARERS (DATA PUBLISHERS)

Data makers often process data collected by a project/activity and package it according to open science principles.

DATA USERS

DATA MAKERS (DATA PROVIDERS)

DATA SHARERS (DATA PUBLISHERS)

Data sharers bear the responsibility of disseminating and building awareness of the data to the public.

Making data open (and FAIR) is a group effort – everybody in the data pipeline has a role to play.

12.10 Lesson 1: Summary

In this lesson, you learned:

- Open data is an essential pillar of open science. Openly sharing data with others enables reproducibility, transparency, validation, reuse, and collaborations.
- Several challenges to creating open data exist, but most have straightforward mitigation measures.
- FAIR principles can be applied to data to make them more open.
- Open-data principles and tasks are used throughout the entire scientific workflow.

12.11 Lesson 1: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/04

Read the statement and decide whether it is true or false.

Open data can be freely used, re-used and redistributed by anyone - subject, at most, to the requirement to attribute and share alike.

- True
- False

Question

02/04

Finish the sentence:

Making data open helps YOU because _____.

- your data can be cited and you will get credit

- you won't lose access to your data, even if you move institutions
- your publications are more likely to get cited when you link to open data
- all of the above

Question

03/04

Choose the FAIR Principles from the list below. Select all that apply.

- Reproducible
- Reusable
- Responsible
- Findable
- Interactive
- Interoperable
- Interspersed
- Accessible
- Authorizable

Question

04/04

Which of the following can help make your data FAIR? Select all that apply.

- Get a license for your data
- Develop your own metadata
- Obtain a PID for your data

13 Lesson 2: Using Open Data

13.1 Navigation

13.2 Overview

In this lesson you learn how to discover, assess, and cite an open data set. You start by exploring repositories and learning about the issues and considerations for searching datasets. You then learn how to determine if the dataset is suitable for your use by learning what to review in documentation, licenses, and file formats. The lesson wraps up with a discussion about the importance of citing the datasets and how to read and follow citation instructions.

13.3 Learning Objectives

After completing this lesson, you should be able to:

- Select data sources and use search techniques to discover open data.
- Assess if a dataset incorporates open access elements that ensure easy reusability.
- Explain the importance of citing open data, and find and follow citation instructions.

13.4 Introduction

Open data isn't always simple to use in your research. Sometimes there are multiple versions of the same dataset, so learning how to discover and assess and then use open data will help you save time.

As an example, look at the monthly average carbon dioxide data from Mauna Loa Observatory in Hawaii. This is a foundational dataset for climate change. Not only is it one of the first observational datasets that clearly showed anthropogenic impacts on the Earth's atmosphere, it constitutes the longest record of direct measurements of carbon dioxide in the atmosphere. These observations were started by C. David Keeling of the Scripps Institution of Oceanography in March of 1958 at a facility of the National Oceanic and Atmospheric Administration [Keeling, 1976].

If you want to make this figure yourself, or use the data for some other purpose, first you will want to find the data. If you search for this dataset, or any data, chances are that you will find a number of different sources. How do you decide which data to use?

If you start with Google and search for "Mauna Loa carbon dioxide data" you will find a lot of results. Here are just some of them:

How do you decide which one to use? In this lesson we will cover how to find, assess relevance, and use open data.

13.5 Discovering Open Data

Open data can be discovered by accessing data repositories, search portals, and publications. A wide variety of these resources are available. A key step is identifying the appropriate search terms for your application. Learning community-specific nomenclature and standards can accelerate your search.

13.5.1 Where to Start Your Search

There are multiple pathways to find research data, and you should be practiced in all of them.

13.5.2 People You Know (Online or In-person!)

When we show up to the present moment with all of our senses, we invite the world to fill us with joy. The pains of the past are behind us. The future has yet to unfold. But the now is full of beauty simply waiting for our attention.

What is the first and best way to find research data? Ask your community, including your research advisor, colleagues, team members, and people online. Knowing where to find reliable, good data is as much a skill and art as any lab technique. You learn this skill set by working with professionals in your field. There is no one source, no one method.

Image source: NASA, Dominic Hart 2023

13.5.3 Publications

Datasets are often attached to scholarly publications in the form of supplementary material. Publication search engines can enable the discovery of relevant publications that you can then use to find data from a particular publication.

13.5.4 Data Search Portals

Data can also be found utilizing a wide variety of search portals including:

Select each tab to find out more information.

GENERIC DATA SEARCH PORTALS

DISCIPLINE-SPECIFIC DATA SEARCH PORTALS

NATIONAL AND INTERNATIONAL DATA SEARCH PORTALS

Generic data search portals enable discovery of a wide variety of data. Not built for specific disciplines, they serve a broader audience. This type of search portal collects and makes data findable. They are not sources of scientific data. These are aggregation services that emphasize quantity, not necessarily quality. This is where citizen scientists often go to find data, and it's a great way for non-professionals to get involved in science.

Examples include:

Google

Kaggle

Wikidata

Open Data Network

Awesome Public Datasets

GENERIC DATA SEARCH PORTALS

DISCIPLINE-SPECIFIC DATA SEARCH PORTALS

NATIONAL AND INTERNATIONAL DATA SEARCH PORTALS

Discipline-specific data search portals enable the discovery of specific types of data. They generally are tailored to meet their community's needs.

Examples include:

NASA Earthdata

CERN

NCBI National Center for Biotechnology Information

EMBL's European Bioinformatics Institute

ISPCR

NOAA Climate Data Online

USGS EarthExplorer

Open Science Data Cloud (OSDC)

NASA Planetary Data System

GENERIC DATA SEARCH PORTALS

DISCIPLINE-SPECIFIC DATA SEARCH PORTALS

NATIONAL AND INTERNATIONAL DATA SEARCH PORTALS

National and international data search portals enable discovery of data produced by or funded by national and international organizations.

Examples include:

US Federal data

EU Data Portal

WHO

The World Bank

data.gov.uk

UNICEF

data.gouv.fr - Open Platform for French Public Data

13.5.5 Repositories

A common way to share and find open data is through data repositories. Many repositories host open data with persistent identifiers, clear licenses and citation guidelines, and standard metadata.

Note that some of our example search portals are also repositories, but not always. Some of the search portals are simply catalogs of information about the data, rather than storage locations for the data themselves.

Select each tab to find out more information.

GENERAL REPOSITORIES

DOMAIN-SPECIFIC REPOSITORIES

INSTITUTIONAL REPOSITORIES

NATIONAL REPOSITORIES

General repositories are not designed for a specific community and are accessible to everyone.

Examples include:

Zenodo

Mendeley Data

Figshare

Dryad

See the Generalist Repository Comparison Chart – a tool for additional repositories and guidance. Dataverse has also published a comparative review of eight data repositories.

GENERAL REPOSITORIES

DOMAIN-SPECIFIC REPOSITORIES

INSTITUTIONAL REPOSITORIES

NATIONAL REPOSITORIES

Specialized repositories (typically for specific data subject matter) provide support and information on required standards for metadata and more.

Some examples are:

Astronomy: Hubble data

Space Biology: NASA GeneLab: Open Science for Life in Space

Space Physics: Heliophysics Data Portal- Solar Space Physics Product Finder (nasa.gov)

GENERAL REPOSITORIES

DOMAIN-SPECIFIC REPOSITORIES

INSTITUTIONAL REPOSITORIES

NATIONAL REPOSITORIES

Many universities and organizations support research data and software management with repositories, known as institutional repositories, to aid their researchers with compliance requirements.

GENERAL REPOSITORIES

DOMAIN-SPECIFIC REPOSITORIES

INSTITUTIONAL REPOSITORIES

NATIONAL REPOSITORIES

National repositories aggregate data and make it available to the public.

Data stored in these repositories are often produced by the government.

Examples include:

https://data.gov/

https://data.europa.eu/en

13.5.6 Challenges with Data Repositories

- Any single repository, search engine or publication search will not have access to all available open data.
- Search terms may not be consistent across sources or fields of science.
- It is essential to become familiar with the standard nomenclatures and appropriate metadata terms for your application.
- There is no sure-fire recipe. You may have to try numerous terms and data sources before finding relevant data.

13.5.7 Activity 2.1: Discovering Open Data

Match the repository type to the correct definition.

General repositories	Designed for all communities and are accessible to everyone
Domain-specific repositories	Repositories that are typically for specific data subject matters
Institutional repositories	Repositories supported by universities and organizations
National repositories	Repositories funded by the government

13.6 Assessing Open Data

Using open data for your project is contingent on a number of factors including quality of data, access and reuse conditions, data findability, and more. A few essential elements that enable you to assess the relevance and usability of datasets include (adapted from the GODAN Action Open Data course):

Practical Questions

- Is the data well described?
- Is the reason the data is collected clear? Is the publisher's use for the data clear?
- Are any other existing uses of the data outlined?
- Is the data accessible?
- Is the data timestamped or up to date?
- Will the data be available for at least a year?
- Will the data be updated regularly?
- Is there a quality control process?

Technical Questions

- Is the data available in a format appropriate for the content?
- Is the data available from a consistent location?
- Is the data well-structured and machine readable?
- Are complex terms and acronyms in the data defined?
- Does the data use a schema or data standard?
- Is there an API available for accessing the data?
- What tools or software are needed to use this data?

Social Questions

- Is there an existing community of users of the data?
- Is the data already relied upon by large numbers of people?
- Is the data officially supported?
- Are service level agreements available for the data?

- It is clear who maintains and can be contacted about the data?

[cite: https://aims.gitbook.io/open-data-mooc/unit-3-using-open-data/lesson-2.2-quality-and-provenance]

Many of these questions may be answered by viewing a dataset's documentation and metadata, as well as a data's format and license, all of which will be discussed further in the next lesson "Making Data Open".

13.7 Using Open Data

13.7.1 The Importance of Citation

Acknowledgements and citations contribute towards fostering a culture of sharing data without fear of ideas or recognition being stolen. If a researcher can trust that their work will be cited, and used to further the development of science, the idea of making data open is more appealing and mutually beneficial. Use of standard citation practices are recommended to ensure due credit is given.

Data citations also aid in the transparency of how data is being used. By citing data, original authors and new researchers can easily track how the data are being used to answer different questions.

13.7.2 Review Citing Guidelines

Many datasets and repositories explain how they'd prefer to be cited. The citation information often includes:

- Authors and their institutions
- Title
- ORCiD
- DOI
- Version
- URL
- Creation date
- Additional fields may also be specified

This is an example of a simple CITATION.cff file. Source: GitHub

---

Most datasets require (at a minimum) that you list the data's producers, name of the archive hosting the data, dataset name, dataset date, and DOI when citing data.

13.7.3 Citing Open Data: Examples

Example from a NASA Distributed Active Archive Center (DAAC)

Matthew Rodell and Hiroko Kato Beaudoing, NASA/GSFC/HSL (08.16.2007), GLDAS CLM Land Surface Model L4 3 Hourly 1.0 x 1.0 degree Subsetted,version 001, Greenbelt, Maryland, USA:Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed on July 12th, 2018 at doi:10.5067/83NO2QDLG6M0

Example from NASA Planetary Data System (PDS)

Justin N. Maki. (2004). MER 1 MARS MICROSCOPIC IMAGER RADIOMETRIC

RDR OPS V1.0 [Data set]. NASA Planetary Data System. https://doi.org/10.17189/1520416

13.8 Lesson 2: Summary

The following are the key takeaways from this lesson:

- Relevant data may be found in a variety of locations and may require some trial and error to find.
- Carefully assess data before using it for your project.
- Data citation is important when using data.

13.9 Lesson 2: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/03

Which of the following methods can be used for data discovery?

- Using appropriate search terms
- Investigating data identified by DOIs in publications
- Identifying relevant data repositories
- All of the above

Question

02/03

Which of the following is/are questions to consider when assessing if a dataset can be used?

- Is the data well described?
- Is the data well-structured and machine readable?
- Is there an existing community of users of the data?
- What tools or software are needed to use this data?
- Will the data be updated regularly?
- Is the publisher's use for the data clear?
- All of the above

Question

03/03

What information is commonly found in a citation file?

- Authors and their institutions
- Title
- ORCiD
- DOI
- Version
- URL
- Creation date
- All of the above

14 Lesson 3: Making Open Data

14.1 Navigation

14.2 Overview

In this lesson, you learn the criteria and tasks needed to ensure that the datasets you make are open and reusable. The lesson starts with a discussion on creating a data management plan and then continues with topics on selecting open data formats and how to include metadata, readme files, and version control for your data. It wraps up with a discussion on open licenses for data.

14.3 Learning Objectives

After completing this lesson, you should be able to:

- Evaluate and select open data formats.
- Add documentation that enables other researchers to assess the relevance of the data. This includes metadata, README files, and version control.
- List two common open licenses used for datasets.

14.4 Planning for Open Data

A best practice when beginning your open data journey is to create a Data Management Plan (DMP). This describes how you will manage, preserve, and release your data during and after a research project. Common elements of DMPs relevant to open data include a description of the following:

What?	Data formats and (where relevant) standards
When?	When and if to share data
Where?	The intended repositories for archived data
How?	How the plan enables reuse of the data
Who?	Roles and responsibilities of the team members in implementing the DMP

In this lesson, we will cover some commonly practiced steps to make data. Specifically, we will focus on the "what" of making data. This covers what data formats you should use and what standards to follow to make the data as open and as readily usable as possible.

As a first step, check if your home institution or funding source has guidance, standards, or templates for DMPs.

14.5 Selecting Data Formats and Tools for Interoperability

14.5.1 Data Format Considerations

Preferred data formats are community supported, machine-readable, non- proprietary, modifiable, and open. It might seem like there are as many data formats as there are different types of data. When you think about selecting a data format, consider the following:

- Is the format compatible with your data type, shape, and size?
- Does the data format have adequate metadata support?
- Are there tools readily available or any specialized tools are required for reading the data format?
- Is the data format routinely used in your field? Community standards ensure compatibility, interoperability, and ease of use when exchanging or sharing data among researchers or organizations of the same community.

Investigate if your funding agency, institutions, and/or data repository has additional requirements for or guidance on data formats.

14.5.2 Non-Open Data Formats

A non-open (unsupported and closed/proprietary) data format refers to a file format that is not freely accessible, standardized, or widely supported by different software applications. Here are some examples of closed/proprietary data formats:

- **Adobe Photoshop (.psd):** The default proprietary file format for Adobe Photoshop, a popular image editing software.
- **Microsoft Word (.doc/.docx):** A proprietary file format used to store word processing data.
- **AutoCAD Drawing (.dwg):** A proprietary data format used for computer-aided design (CAD).

Software applications that can read but not create DOC, PSD, or DWG formatted data usually do not fully support all the features, layers, specifications, and inner workings of the original file.

Some challenges of using data in non-open formats include:

- Trouble opening the file due to compatibility issues.
- The need to install additional software or converters, leading to frustration and inconvenience.
- Initial setback dampens the enthusiasm for using your data.
- Converting the data to a universal format can lead to unique formatting or features that do not translate well, making the data lose part of its value.
- New open-data policies can limit the sharing of proprietary data as it is often non-compatible with the concept of easy distribution.

14.5.3 Open Data Format Examples

Some examples of open data formats include:

Select each card to find out more information.

Comma Separated Values (CSV)	For simplicity, readability, compatibility, easy data exchange.
Hierarchical Data Format (HDF)	For efficient storing and retrieving data, compression, multi-dimensional support.
Network Common Data Form (NetCDF)	For self-describing and portability, efficient data subsetting (extract specific portions of large datasets), standardization and interoperability.

Investigation-Study- Assay (ISA) model for life science studies	For structured data organization, data integration and interoperability among experiments, reproducibility and transparency.
Flexible Image Transport System (FITS)	As a standard for astronomical data, flexible and extensible metadata and image headers, efficient data compression and archiving of large datasets.
Common Data Format (CDF)	For self-describing format readable across multiple operating systems, programming languages, and software environments, multidimensional data, and metadata inclusion.

By embracing open standards, authors can avoid unnecessary barriers and maximize their chances of making data useful to their communities.

14.6 Making the Data Reusable Through Documentation

14.6.1 Adding Documentation and Metadata for Reusability

Metadata and data documentation describe data so that we and others can use and better understand data. While metadata and documentation are related, there is an important distinction. Metadata are structured, standardized, and machine readable. Documentation is unstructured and can be any format (often a text file that accompanies the data).

To better understand documentation and metadata, let's take an example of an online recipe. Many online recipes start with a long description and history of the recipe, and perhaps cooking or baking tips for the dish, before listing ingredients and step-by-step cooking instructions.

- The ingredients and instructions are like metadata. They can be indexed and searched via Google and other search engines.
- The descriptive text that includes background and context for the recipe are like documentation. They are more free-form, and not standardized.

We already discussed metadata earlier in this module, but it's important enough that we will repeat ourselves a little bit! We will also discuss other types of documentation, like README files.

14.6.2 Metadata: for Humans and Machines

Metadata can facilitate the assessment of dataset quality and data sharing by answering key questions. It is also the primary way users will find information about your dataset. It includes key information on topics, such as:

- How data were collected and processed
- What variables/parameters are included in the dataset
- What variables are included and what variables are related to
- Who collected the data (science team, organization, etc.)
- How and where to find the data (e.g., DOI)
- How to cite the data
- Which spatio-temporal region/time the data covers
- Any legal, guideline, or standard information about the data

14.6.3 Why Add Metadata?

Metadata enhances searchability and findability of the data by potentially allowing both humans and machines to read and interpret datasets. Benefits to creating metadata about your data include:

- Helps users understand what the data are and if/how they can use/cite it.
- Helps users find the data, particularly when metadata is machine- readable and standardized.
- Can make analysis easier with software tools that interpret standardized metadata (e.g. Xarray).

To be machine readable, the metadata needs to be standardized. See an example of a community-accepted standard for labeling climate datasets with the CF Conventions.

There are also software packages that can read metadata and enhance the user experience significantly as a result. For instance, Xarray is an open-source, community developed software package that is widely used in the climate and biomedical fields, among many others. According to their website, "Xarray makes working with labeled multi-dimensional arrays in Python simple, efficient, and fun!". It's the "labeled" part where standardized metadata comes in! Xarray can interpret variable and dimension names without user input, making the workflow easier and less prone to making mistakes (e.g. users don't have to remember which axis is "time" - they just need to call the axis with the label "time").

Many standards exist for metadata fields and structure to describe general data information. Use a standard from your domain when applicable, or one that is requested by your data repository.

14.6.4 Metadata Tagging Best Practices

Useful and informative metadata:

- Uses standards that are commonly used in your field.
- Complies with FAIR Principles.
- Is as descriptive as possible.
- Is self-describing.

Remember, the more metadata you add, the easier it will be for users of your data to use it effectively. When in doubt:

- Seek and comply with repository/community standards.
- Investigate open science online resources for metadata, e.g., Turing Way.

14.6.5 Accompanying Documentation

When creating your data, in addition to adding metadata, it is a best practice to create a document that users can refer to. The document can be done as a README file, a user guide, or even a quick start (or all three).

README and other documentation files can include information such as:

- Contact information
- Information about variables
- Information about uncertainty
- Data collection methods
- Versioning and license references
- Information about the structure and file naming of the data
- References to publications that describe the dataset and/or it's processing

The intent is to help users quickly understand how they might use the data and to answer any commonly asked questions about your data. You can read more information and view a README template along with an example (particularly relevant for the medical sciences) at this Harvard Medical School website.

14.6.6 Data Versioning Guidelines

Establish a versioning schema for your data. This is a method for keeping track of iterations of data that features track changes and the ability to revert to a previous revision.

Proper versioning generates a changed copy of a data object that is uniquely labeled with a version number. This enables users to track changes and correct errors.

Proper versioning preserves data quality and provenance (the origin, history, and processing steps that lead to the dataset) by:

- Providing a record of traceability from the data's source through all aspects of its transmission, storage, and processing to its final form.
- Saving data files at key steps along the way.
- Aiming for downstream verification/validation of original findings.

14.7 Making the Data Reusable Through Licensing

Image source: xkcd.com

Data is the intellectual property of the researcher(s), or possibly of their funder(s) or supporting institution(s). Data is intellectual property, but that does not mean it cannot be used by other researchers (with appropriate attribution).

"By applying a license to your work, you make clear what others can do with the things you're sharing, and also the conditions under which you're providing them (like cite you). You can also require others who copy your work to do things in return."

Open Science Knowledge Base

If you don't license your work, others can't/shouldn't re-use it - even if you want them to. As mentioned previously in this module, a license is a legal document that tells users how they can use the dataset. It is important to understand the licensing conditions of a dataset before data reuse to avoid any copyright infringement or other intellectual property issues.

A dataset without a license does not mean that the data is open; using a licenseless dataset is not ethical. Contacting the data creator and getting explicit permission, while suggesting they apply a license, is the best path forward.

Understanding when and where the license applies is crucial. For example, data created using US Government public research funds is, by default, in the public domain. However, that only applies to the jurisdiction of the United States. In order for this to apply internationally, data creators need to select an open license.

There are several different types of licenses that build on each other. Creative Commons (CC) licenses are often used for datasets. CC0 (also known as "public domain") is the license that allows for the most reuse because it has the least restrictions on what users can do with it.

Although the CC0 license does not explicitly require citation, you should still follow community best practices and cite the data source. CC-BY is another common license used for scientific data that requires citation. From there, you can add restrictions around commercial use, ability to adapt or modify the data, or requirements to share with the same license. These other flavors all reduce usability by adding restrictions, such that other scientists may be unable to use the data because of institutional or legal restrictions. Funding agencies may require use of a specific license. For public agencies, this is often CC-0 or CC-BY, to maximize their return on investment and ensure widest possible re-use.

14.7.1 Example Data Licenses and Reuse

Here is an example of how a data license can affect reuse. Coupled Model Intercomparison Project Phase 6 (CMIP6) consists of the "runs" from around 100 distinct climate models being produced across 49 different modeling groups. This is the data that is used to understand what our future climate might look like. You have probably seen images that use this data in articles about Earth's changing climate and how it may impact our lives. Previous versions of these data were licensed CC-BY-NC-SA (cite-noncommercial- sharealike).

Figure citation: IPCC "Framing and Context in : In: Global warming of 1.5°C. An IPCC Special Report" 2020

This meant that any commercial use was restricted. Insurance companies, global corporations, and any type of organization that wanted to use them for commercial use - were having to do their own modeling or just deciding to not develop resources related to climate projections (such as fire risk, flooding risk, and how that may affect transportation, commerce, and where we live). This directly impacted the reuse of this data and created additional work. The latest version of CMIP data is moving to CC-BY because of the negative impacts from the -NC-SA restrictions.

14.8 Lesson 3: Summary

Following are the key takeaways from this lesson:

- It is best practice to create an open data management plan that includes open-related topics.
- A critical step to making open data is evaluating and selecting open data formats.
- Always add documentation that enables other researchers to assess the relevance and reusability of your product. This includes metadata, README files, and version control details.
- It is important to assign an open license to your data to enable reuse.

14.9 Lesson 3: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/04

Which of the following are steps you should take when making a Data Management Plan?

- Evaluate different data formats
- Test your metadata for compliance
- Create a small collection of test data

Question

02/04

Which of the following are considerations when choosing a file format?

- The format has adequate metadata support
- Tools are readily available to read the data format
- The data format is widely used in your community
- All of the above

Question

03/04

Read the statement below and decide whether it's true or false.

Metadata is only useful for using data in interoperable tools and does not enhance searchability and findability of data.

- True
- False

Question

04/04

Read the statement below and decide whether it's true or false.

When a dataset does not explicitly require citation, such as the CC0 license, it is still recommended that you cite the data source.

- True
- False

15 Lesson 4: Sharing Open Data

15.1 Navigation

15.2 Overview

In this lesson, you learn about the practice of sharing your data. The discussion starts with a review of the sharing process and how to evaluate if your data are sharable. Next, you take a look at ensuring your data is accessible with a closer look at repositories and the lifecycle of data accessibility from the selecting a repository to maintaining and archiving your data. The lesson then discusses some steps to make the data as reusable as possible, and concludes with a section about considering who will help with the data sharing process.

15.3 Learning Objectives

After completing this lesson, you should be able to:

- Recognize institutional variables, issues of security, and timing that affect your decision to share data.
- Recall the features, inherent responsibilities, funding considerations, and sponsor requirements that researchers should consider when selecting a repository to share data.
- Describe the tools and list some best practices that optimize the shareability of data.

15.4 Data Sharing Process Overview

Sharing data is a critical part of increasing reproducibility of results. Whether it's new data we collect ourselves or data that we process in order to do our analysis, we end up sharing some form of data. We need to think about what data we will share and how to best ensure that it will be open and usable by others.

Data sharing should typically be done through a long-term data center or repository which will be responsible for ingesting, curating, and distributing/publishing your open data. You are responsible for providing information/metadata to help make your data be readily discoverable, accessible, and citable. The cost of archiving and publishing data should also be considered.

15.4.1 So You Want to Share Your Data

Once you have decided to share your data, there are a number of questions you will have to answer to help you plan and that should be included in your data management plan (DMP):

What?	Data formats and (where relevant) standards
When?	When and if to share data
Where?	The intended repositories for archived data
How?	How the plan enables reuse of the data
Who?	Roles and responsibilities of the team members in implementing the DMP

In this lesson, we will be covering some steps toward making data. Specifically, we will focus on the "when", "where", "how", and "who" sections of a DMP.

15.4.2 Open Data Sharing Process

In general, sharing your open data requires the following steps:

1. Make sure your data can be shared
2. Select or identify a repository to host your data
3. Work with your repository to follow their process and meet their requirements
4. Make sure your data is findable and accessible through the repository and is maintained and archived
5. Request a DOI for your data set so that it is easily citable
6. Choose a data license

Sometimes, you may be able to work with a well-staffed repository that will handle many of these steps for you. Otherwise, it is your responsibility to follow the above steps to share your data openly.

15.5 When and If to Share Data

15.5.1 When to Share Data?

The decision of when to share should be discussed with everyone on the team and documented in the data management plan. Funding agencies and organizations may have specific requirements about when data must be shared, but here we encourage you to think about whether it is feasible or possible to share even earlier than required by your funder. There are different times when data could be shared:

- Advanced Sharing: Sharing at the time of collection, or soon after. Some funding agencies require this, or allow for a short 'embargo' period, but a reason (quality checks, calibrations, etc.) is usually required. This maximizes data reuse and impact and can result in increased collaborations.
- Intermediate Sharing: At the time of publication. Many publications (and some funding agencies) require sharing data that is needed to reproduce results at the time of publication.
- Minimum Sharing: End of grant. All scientifically useful data should be shared by the end of the research grant.
- No Sharing: There are many reasons data should either be restricted or not shared at all.

As discussed previously in this curriculum, there are many benefits to sharing as early as possible. Early (advanced) sharing can lead to new and unexpected discoveries and expand your collaboration network. Remember, that even when you share data, you are still the world expert on that data! So often, when people want to work with the data, they will reach out to you to collaborate.

15.5.2 Should the Data be Shared?

Before datasets are shared, it's important to consider any restrictions to your permission to share and ensure that your contributors – including sample and data donors – approve its release.

Data should be as open as possible and as closed as necessary.

- Opening our data is a powerful way to enable discovery, transparency, and scientific progress.
- Some data are subject to laws, regulations, and policies which limit the release of the data.
- Your local institution may have additional policies and resources – investigate them early and often.

15.5.3 Verify Your Data is Sharable

Before you decide where to share your data, you must make sure you can share your data.

Data needs to be as open as possible and as closed as necessary…

- Open data is a powerful way to enable discovery, transparency, and scientific progress
- But, some data are subject to laws, regulations, and policies which limit the release of the data
- Your local institution may have additional policies and resources – investigate them early and often

Specific considerations that might prevent the sharing of your data include:

- A country's military secrets or violations of national interests
- Private medical information or an individual's personal data
- Indigenous/cultural/conservation concerns
- Intellectual Property, Patented
- Other - please think about what you are sharing and the implications of sharing it (for example - do you have permission from everyone involved?)

In the first module of this curriculum, we listed several reasons why certain research products should not be shared. We will review some of these reasons, and go into more detail on a few that are particularly relevant to data.

15.5.4 Export and Security Considerations

Relevant laws and regulations that may prevent the release of data include but are not limited to:

- International Traffic in Arms Regulation (ITAR), which regulates the manufacture, sale, distribution, and export of defense-related articles and services.
- Export Administration Regulations (EAR), which regulates the manufacture, sale, distribution, and export of commercial and dual- use items, technology, and information not already covered by ITAR.

Example: NASA Space System Protection Standard

NASA STD 1006.1 Space System Protection Standard, which establishes protection requirements to ensure NASA missions are resilient to purposeful threats.

15.5.5 Controlled Information Considerations

Some regulations and policies that may prevent the sharing of data include but are not limited to:

- Health Insurance Portability and Accountability Act (HIPAA), which established standards to protect sensitive patient health information from disclosure.
- Controlled Unclassified Information provides standards for handling unclassified information that requires safeguarding or dissemination controls consistent with laws, federal regulations, and policies.
- Federal laws and regulations governing classified information or security requirements.

15.5.6 Intellectual Property Considerations

Data may be subject to intellectual property, copyright, and licensing concerns. A few of the relevant regulations and policies include patent or intellectual property laws including the Bayh-Dole Act, which enables universities, nonprofit research institutions, and small businesses to own, patent, and commercialize inventions developed under federally funded research programs.

Example: NASA FAR Supplement 1852.227

NASA FAR Supplement 1852.227, which outlines patent and data rights for government contracts.

Many research institutions have resident experts in intellectual property, copyright, and patent law. They can be a great resource if you have any questions or concerns.

15.6 Where to Share Data

Data can be shared in a variety of locations. While sharing data via email or websites is popular, they are not recommended as they do not meet the requirements for findability or long-term archival support. Sharing data as part of the supplemental material of a peer reviewed publication, especially for small data sets, is acceptable in some fields. A long term repository that provides a permanent identifier is the best option for sharing of data.

15.6.1 Selecting a Data Repository

If you do not already have a data repository in mind, consider the following to narrow down your options:

- Does your funding sponsor require a specific data repository?
- Does your organization/institution recommend a specific data repository?
- Is there a domain-specific repository that is widely-used in your research field?
- Does the repository provide open data access?
- Do you think the tools offered by the repository for data discovery and distribution are suitable for your data and FAIR?
- Does the repository require funding from your project, does it fit within your budget and does it require sustained support beyond the project life cycle?

Find and compare the services, benefits and limitations of the repositories you are considering. Each repository will have its own processes and requirements for accepting and hosting your data depending on their level of funding, purpose, and user base.

Similarly, each repository will provide a different set of functionality and services depending on their level of funding, purpose, and user base.

Data with privacy concerns may have additional anonymization or approval processes or restrictions on who can access the data.

A good overview of desirable characteristics presented by the White House is given here.

15.6.2 Ensuring Accessibility

Good repositories will share (or offer) your open data through standard protocols, like HTTPS or SFTP. Common ways to do this are:

- Allowing users the ability to see a list of files that they can click and download via an intuitive interface.
- Creating a documented API for users to generate a list of file links that meet search criteria that they can download in an automated fashion (i.e., machine-to-machine data access).

Additionally, repositories can require authorization and authentication (e.g., logins with usernames/passwords) to access data. While this is allowed under FAIR principles, it may violate Open Science principles if not everyone is able to obtain a login.

15.6.3 Working with a Repository

START WORKING WITH A REPOSITORY

MAINTAINING DATA AT A REPOSITORY

ARCHIVING DATA AT A REPOSITORY

Repository requirements can vary widely. Always review a repository's requirements to see what actions you need to take once you're ready to start working with them. Also note that some repositories have staff that will help with the process of sharing data, while others rely on the user to know how to share their own data.

If you use a repository that has staff to help you with the process, they may want to review and comment on your data management plan.

The repository may request that you produce some test of sample data in order to assess:

That the data format you intend to use is supported.

That data variables are named as expected.

That metadata vocabulary is correct.

That repository-specific requirements are met.

This conformity check can identify misunderstandings early and result in a smooth final submission of your data to the repository.

START WORKING WITH A REPOSITORY

MAINTAINING DATA AT A REPOSITORY

ARCHIVING DATA AT A REPOSITORY

As you progress through your project lifecycle, utilize your repository's update, revision and resubmission processes to keep the archived data products up to date. Any new versions of the data you want to share through the repository will need to go through a similar process as your initial data set.

Any new versions of the data you want to share through the repository should go through the same DMP review, compliance check, and upload procedure as your initial data set.

START WORKING WITH A REPOSITORY

MAINTAINING DATA AT A REPOSITORY

ARCHIVING DATA AT A REPOSITORY

When your project ends, ensure you've updated and uploaded any companion documentation (discussed in the previous lesson "Making Open Data") with your final version (even if only a single version of the data was made).

Make sure the repository will keep your data (or at least your metadata) on- line for a reasonable period of time after your project ends.

If any data issues are found after the conclusion of your project, make sure the repository will still accept data revisions, if they are needed.

15.7 How to Enable Reuse of Data

15.7.1 Obtaining a DOI

Individuals cannot typically request a DOI (digital object identifier) themselves but rather have to go through an authorized organization that can submit the request, such as:

- The data repository
- Your organization
- The publisher (if the data set is part of a publication)

Data makers should provide summary information for DOI landing page(s) if required. Data sharers should accommodate data providers' suggestions and comply with DOI guidelines and create landing page(s). If possible, reserve a DOI for you ahead of creating your data.

15.7.2 Ensuring Findability

Repositories handle the sharing, distribution, and curation of data. Additional services they may provide include:

- The assignment of a persistent identifier (like a DOI) to your data set
- The indexing and/or registration of your data and metadata in various services so that they can be searched and found online (i.e., through search engines).
- The provision of feedback to data makers to help them optimize their metadata for findability.
- Coordinating with data makers to ensure metadata refers to the DOI.
- Ensuring the DOI is associated with a landing page with information about your data.

15.7.3 Making it Easy to Cite Your Data

The goal is to make it easy to cite your data. Best practices include:

- Include a citation statement that includes your DOI.
- Different repositories and journals have different standards for how to cite data. If your repository encourages it, include a .CFF file with your data that explains how to cite your data.

- Clearly identify the data creators and/or their institution in your citation.
  - This allows users to follow up with the creators if they have questions or discover issues.
  - Include ORCiD of data authors where possible in the citation.

Now that your data are at a repository and have a citation statement and DOI, publicize it to your users and remind them to cite your data in their work!

15.8 Who is Responsible for Sharing Data

Sharing data openly is a team effort. An important part of planning for open data is planning and agreeing to roles and responsibilities of who will ensure implementation of the plan.

So what needs to be done? Documenting these roles and responsibilities in your Data Management Plan will help your team stay organized and do science faster! A well-written, detailed plan should include:

15.8.1 Who Will Move Data to a Repository

Once you are ready to send your data to your repository, find the repository's recommendations for uploading data. Determine who will work with your repository to accomplish the following types of activities:

- Provide information on data volume, number of files, and nature (e.g., revised files)
- Check that the file name follows best practices
- How will the data be moved? (especially when files are large)
- Check the data! Verify the integrity of the data, metadata, and documentation transfer

15.8.2 Who Will Develop the Data Documentation and Metadata

Determine who will work with your repository, inventory the transferred data, metadata, and documentation. This role might include the task of populating any required metadata in databases to make the data findable.

You may be able to accomplish some of these tasks through a repository's interface. However, some types of repositories may require you to interact with their administration teams. For this role, determine who will:

- Provide suggestions to organize data content and logistics-
- Develop the metadata
- Develop the documentation (e.g., README file or report)

- Extract metadata from data files, metadata files (if applicable), and documentation to populate the metadata database and request additional metadata as necessary

15.8.3 Who Will Help With Data Reuse

Once the repository has made your data available, someone from your team must test access to the data (its accessibility) and distribution methods (its findability). If possible, identify who will work with your repository to optimize/modify tools for intuitive human access and standardize machine access. This role requires someone who to:

- Clearly communicate the open protocols needed for the data/metadata.
- Provide actual data use cases to data publisher to optimize/modify data distribution tools based on available metadata.
- Understand the access protocol(s) and evaluate implications to targeted communities and user communities at large in terms of accessibility.

15.8.4 Who Will Develop Guidance on Privacy and Cultural Sensitivity of Data

Sharing data should be respectful of the communities that may be involved. This means thinking about privacy issues and cultural sensitivities. Who on your team will identify and develop guidance on:

- Privacy concerns and approval processes for release - is the data appropriately anonymized?
- How to engage with communities that data may be about.
- How data can be correctly interpreted.
- Are there any data restrictions that may be necessary to ensure the sharing is respectful of the community the data involves, eg. collective and individual rights to free, prior, and informed consent in the collection and use of such data, including the development of data policies and protocols for collection?

15.9 Lesson 4: Summary

The following are the key takeaways from this lesson:

- When and if to share data? Determine at what point in a project it makes the most sense to share our data. Remember, not all data can or should be shared.
- Where to share data? Sharing in a public data repository is recommended, and there are many types of repositories to choose from.
- How to enable reuse? Ensure appropriate, community-accepted metadata, assign a DOI, and develop a citation statement to make sure it can be easily found and cited.

- Who helps share data? There are many steps in making and sharing data and it's important to think about who will be responsible for each step.

15.10 Lesson 4: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/04

Data cannot be shared if it is:

- ITAR controlled
- Controlled Unclassified Information
- Subject to intellectual property, copyright, and licensing concerns
- All of the above

Question

02/04

Select the option you think is correct to complete the sentence.

It is best practice to start working with a repository _____.

- As early as possible
- When you have test data ready
- After you obtain a DOI
- When you are ready to release your data

Question

03/04

Which one of the following might be able to help you get a DOI for your data:

- The repository you are working with
- Your home organization
- A journal you are submitting a manuscript and data to
- All of the above

Question

04/04

Which of the following are roles to consider when sharing data? Select all that apply.

- Develop guidance on privacy and cultural sensitivity of the data
- Develop the data documentation and metadata
- Assign the data a DOI
- Verify the integrity of the data, metadata, and documentation transfer

16 Lesson 5: From Theory to Practice

16.1 Navigation

16.2 Overview

In this lesson, you will get some practice writing a data management plan. You will then learn how you can get involved in open data communities. You will also learn about resources you can start to use and training you can take to start your journey with open data.

16.3 Learning Objectives

After completing this lesson, you should be able to:

- Describe the steps toward writing a data management plan.
- List opportunities for involvement in the open data communities.
- Identify additional open science resources and list ways to continue training.

16.4 Writing an Open Science and Data Management Plan

The process, responsibilities, and factors to consider when creating an open science and data management plan have been presented throughout this module. Common elements of DMPs relevant to open data include a description of the following:

What?	Data formats and (where relevant) standards
When?	When and if to share data
Where?	The intended repositories for archived data
How?	How the plan enables reuse of the data
Who?	Roles and responsibilities of the team members in implementing the DMP

Two great places to start are https://dmptool.org/ and https://dmponline.dcc.ac.uk/. You will need to create a free login to use these tools, but both websites walk researchers through the steps of writing a DMP. There are even some existing DMP templates stored within DMP Tool that may be targeted towards a funding body you are interested in.

There are also public examples of data management plans at

https://dmptool.org/public_plans and

https://dmponline.dcc.ac.uk/public_plans.

If you are applying for funding, it is almost guaranteed that there will be specific requirements detailed in the funding opportunity. For example, the funder may require a certain license or use of a specific repository. Make sure to cross reference your plan with these requirements!

16.4.1 Activity 5.1: Review a data management plan

Take a look at the example of a public data management plan from the ThEBES project.

In the DMP, answer the following questions:

1. What: Data formats and (where relevant) standards.
2. When: When and if to share data.
3. Where: The intended repositories for archived data.
4. How: How the plan enables reuse of the data.
5. Who: Roles and responsibilities of the team members in implementing the DMP.

16.5 Open Data Communities and You

16.5.1 Getting Involved with Open Data Communities

There are numerous ways to get involved with and support open data communities, including starting your own community.

REPOSITORIES

STANDARDS COMMITTEES

CONFERENCES , WORKSHOPS AND SPECIAL SESSIONS

Contribute to open data repositories

Many repositories have user committees to provide them with advice and feedback (and are often looking for volunteers to serve)

Subscribe to repository mailing lists and social media accounts

REPOSITORIES

STANDARDS COMMITTEES

CONFERENCES , WORKSHOPS AND SPECIAL SESSIONS

Volunteer to serve on a standards committee

Provide input to the standards committee

Subscribe to the mailing lists focused on standards

REPOSITORIES

STANDARDS COMMITTEES

CONFERENCES , WORKSHOPS AND SPECIAL SESSIONS

Organize a gathering around open data

Participate in a gathering around open data

16.6 Additional Resources

16.6.1 Resources for More Information

In addition to the resources listed elsewhere in this training, the below community resources are excellent sources of information about Open Data.

References and Guides:

NASA Science Mission Directorate's Open-Source Science Guidance for researchers

CLICK TO LEARN

Open Data Module from OpenSciency

CLICK TO LEARN

Resources for open data through the US federal government

CLICK TO LEARN

The FAIR Principles from FAIRsharing.org

CLICK TO LEARN

The Open Data Handbook

CLICK TO LEARN

Reproducible Research and Data Analysis from FOSTER Open Science

CLICK TO LEARN

Data publishing guidelines from the Data Management Expert Guide

CLICK TO LEARN

DataTags: A Harvard University Privacy Tools project to help researchers use and share sensitive data

CLICK TO LEARN

Best Practices for Data Publication in the Astronomical Literature by Chen et al.

CLICK TO LEARN

16.6.2 Opportunities for More Training About Open Data

In addition to the resources listed elsewhere in this training, the community resources listed below provide excellent information on Open Data.

Additional training:

GODAN MOOC about how to use, make, and share open data

CLICK TO LEARN

Data literacy lessons for an array of disciplines from the Carpentries

CLICK TO LEARN

MOOC on Data Management, Sharing and Services for Agriculture Development

CLICK TO LEARN

16.7 Lesson 5: Summary

Now that you have completed the lesson, you should be able to start your journey with open data:

- You now know the steps and have practice writing a sample data management plan.
- There are a variety of ways to get involved in the open data community.
- There are numerous resources available to get more information and take more training about open data.

16.8 Lesson 5: Knowledge Check

Question

01/02

Which of the following are steps to include in a data management plan (DMP)? Select all that apply.

- What: Data formats and (where relevant) standards
- When: When and if to share data
- Why: Why you used an online DMP template
- Where: The intended repositories for archived data
- Which: Which data communities you are involved in
- How: How the plan enables reuse of the data
- Who: Roles and responsibilities of the team members in implementing the DMP

Question

02/02

What are the three broad categories of getting involved with open data communities discussed in this lesson? Select all that apply.

- Engage with repositories
- Join standards committees
- Write a data management plan (DMP)
- Volunteer and participate in conferences, workshops, and special sessions
- Identify data sharing roles and responsibilities on your team

16.9 Open Data Summary

Congratulations! Now that you have completed the module, you should be able to do the following:

- Explain what open data means, its benefits, and how FAIR principles are used.
- Discover open data, assess the data for reuse by evaluating provided documentation, and cite the data as instructed.
- Create an open data management plan, select open data formats, add the needed documentation, including metadata, readme files and version control, to make the data reusable and findable.
- Evaluate whether your data should and can be shared, and use the data accessibility process, including adding a DOI and citation instructions to enable it to be findable and citable.

Part IV

OS101 Module 4: Open Code

Welcome to Open Science 101: Open Code

About This Module

This module focuses on the practice and application of open code as part of the open science workflow. It provides a 'how to' process that follows the code development lifecycle and "Use, Make, Share" framework. Some of the key topics discussed include: benefits and limitations of open code, how to discover and assess code, considerations and methods for programming following open principles, and finally when and how to share your code.

Module Learning Objectives

After completing this module, you should be able to:

- Explain what open-source software means, including the software development cycle, the benefits, some common limitations, and how they are addressed.
- Assess open-source software for reuse by evaluating provided documentation, including README files and licensing details, and then cite the software appropriately.
- Create an open-source software management plan that includes the strategy for selecting open software dependencies and open repositories, and how open elements including metadata, README files and version control, will be included to make the software reusable and findable.
- Evaluate whether your open-source software can be shared and the best options for sharing to increase visibility.
- List the responsibilities a software developer has once the open-source software is shared including managing legal requirements and ensuring the software is maintained.

Key Terms

Select the term to see the description.

Source Code – Human-readable set of statements written in a programming language that together compose software. Programmers write software in source code, often saved as a text file on a computer. The terms code and source code are often used interchangeably.

Software – This general term is used for computer programs and applications that provide users some degree of utility or produce a result or service. Software can be distributed in executable form, as source code, or as a service via the internet.

Software License – A document that states the rights of the developer and user of a piece of software.

Open-Source License – A software license, approved by the Open Source Initiative (OSI) as compliant with the Open Source Definition, granting permissions for anyone to inspect, use, modify, and distribute the software's source code for any purpose. Similar standards may be promulgated by other organizations.

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. Synonymous with open code.

Closed-Source Software – Proprietary software with source code that is not publicly available. Only the original authors, not the users, have rights to copy, modify, update, and edit the source code. Users don't have access to code.

Derivative Work – A creative work that is derived from or based upon a preexisting creative work and in which the preexisting work is translated, altered, arranged, or transformed in a manner that requires permission from the copyright owner of the original work or are from works in the public domain.

Version Control – A system to automatically manage changing versions of a computer file, especially one that contains source code. In software development, version control preserves a complete history of changes to the source code and enables a developer to roll back to an earlier version if needed.

Code Repository – A central storage location for the source code. Code repositories may contain source code in one or more programming languages. Repositories may provide tools for merging inputs from developers, automated testing to verify the proper functioning of source code, version control to track changes over time, and project management features. These sites may not promise long-term retention.

Software Repository – Online collections of stand-alone applications or software packages. Repositories typically control access and track the deployments and downloads of packages. These may include source code or executables.

Long-Term Repository – A service for long-term retention and referencing historical and contemporary software. Repositories facilitate sustainable citation of code.

Definitions credit: National Academies of Sciences, Engineering, and Medicine. 2018. Open-source Software Policy Options for NASA Earth and Space Sciences. Washington, DC: The National Academies Press. https://doi.org/10.17226/25217 and other resources.

Navigation

Lesson 1: Introduction to Open Code

- Overview

Lesson 2: Using Open Code

Lesson 3: Making Open Code

Lesson 4: Sharing Open Code

Lesson 5: From Theory to Practice

17 Lesson 1: Introduction to Open Code

17.1 Navigation

17.2 Overview

This lesson defines the key terms, core principles, benefits, and challenges of open code. The practice of making code openly available to the public occurs within a spectrum from more to less protected. Ethical and legal conditions can limit the degree of openness that researchers can permit. This lesson will introduce the critical questions to consider when determining the appropriate accessibility of code to external users along with best practices to overcome common constraints to maximize availability. The lesson concludes with a discussion on the software lifecycle and how it fits with the "Use, Make, Share" framework and its relationship to a management plan.

17.3 Learning Objectives

After completing this lesson, you should be able to:

- Define open-source software and distinguish it from closed-source software.
- List common benefits and challenges to the production of open code and describe how researchers can respond to some of the challenges while maximizing openness when appropriate.

- Describe the function and purpose of a Software Management Plan, as its utility as a guidebook for everyone involved in a scientific project.

17.4 Success Stories

Why does good science demand that researchers make their code open-access? Sharing your code (and data) makes it easier for others to reproduce your results, helping to validate findings and reduce resources required to duplicate experiments. As a bonus, this decision can lead to new collaborations made possible through a shared dataset and common understanding of scientific material.

Many journals and funding agencies require that you share your code at the time of publication. However, the prospect of opening code up to criticism, not receiving attribution, or missing out on a result that external researchers discover can deter scientists from making their code open-access. What if people find an error? What if they criticize your coding style? What if they take your code and publish a new result without including you? This module will help you gain confidence in sharing your code by walking you through the basic details to consider when practicing open-science.

Let's review some well-known examples of groups that shared their code and what the impacts were:

Use buttons to navigate between the examples.

The first image of a black hole would not have been possible in this decade if all the required code had to be written solely by the scientists involved. These scientists were able to use well-tested, community accepted open-source software to conduct their analysis and create this now famous image. Dr. Katie Bouman and her team commended the critical role that open-source contributors played in her team's effort to image the first black hole. This breakthrough was made possible by open-source libraries that provided robust and freely available code. The code used to capture this image was crafted by 21,485 contributors. Sophisticated iterative data processing pipelines and algorithms used by Dr. Bouman's team were community developed and tested, making robust and reproducible science possible without having to rewrite every piece of software needed.

This is the Ingenuity helicopter, or as the engineers call her, Ginny. She got to Mars by hitching a ride on the Perseverance rover, landing in the Jezero Crater in 2021.

This is a video of Ginny's first flight. She took off, got about ten feet off the ground, did a spin and landed. This groundbreaking flight proved that powered flight on Mars is achievable, opening up the door for an entirely new era of exploration.

But Ginny's achievements also reflect another new era; one of truly open and inclusive science.

Behind that 4-pound helicopter are more than 12,000 people who contributed code, documentation, design, and more thanks to the open-source software which was used to power her. Everyone who contributed to the open-source software libraries that Ginny used received a badge on their GitHub page that showed they helped fly the first helicopter on Mars.

In addition, Ginny's final software developed at the Jet Propulsion Lab, called F prime, was itself open-source and has been used since in flight research, drones, and CubeSats. In fact, F prime had been copied to other people's repositories more than 1,200 times.

Most space telescope data is embargoed for 12 months with only the lead scientist and their selected team allowed to work with the data. In a unique case, a small portion of data from NASA's new James Webb Space Telescope (JWST) offered an early-release program. This JWST data was made available immediately.

How scary is that? To know that everyone you know is going to have access at the exact same time. The anxiety and stress of feeling as if you don't publish first, you might not have a job, or you might not have the next job that you want.

In one case, a team decided to work fully in the open and collaborate with this early- release data. The result? 20+ planned papers and the first discovery of carbon dioxide on another planet - hinting at the possibility of discovering new life.

Co-author Dr. Natasha Batalha employed open science principles to enable this rapid discovery using the new JWST data. In the years leading up to the JWST release, Dr. Batalha's team formed a collaborative group of 341 members. Once JWST data was made public, the data reduction and scientific interpretation could be reproduced through open software then archived. The research team's first article was made available as open- access on an archived preprint server and published in Nature.

Notably, Dr. Batalha's team published the first identification of CO2 in an exoplanet's atmosphere from spectra taken with JWST. This was conducted with JWST's Early Release Science Program data, the first science data taken by the facility. The team worked in an open-format from ideation, to analysis, through to publication and communication.

This example illustrates the benefits of applying open science principles to rapidly produce meaningful research. The team worked in an open format from ideation, to analysis, through to publication and communication.

New open-source sets of climate models incorporate features that aim to make climate research more collaborative, efficient and reliable.

Scientists have published an open-source framework of climate models (Isca) which contains models that are easy to obtain, completely free, documented, and come with software to make installation and operation easier. All changes are documented and can be reverted. Therefore, anyone can easily use the same models.

Although the Isca model was initially used to examine the tropical upper atmosphere, researchers from other fields of science have used it to study the life cycle of weather systems, the Indian monsoon, and the effect of volcanic eruptions on climate.

New research across all of these fields was possible within only one year of the Isca's first publication. This is how we want all of science to work!

Credit:

https://theconversation.com/making-climate-models-open-source-makes-them-even-more-useful-90929

17.5 Definitions and Considerations of Open Code

All science builds on what has already been accomplished. Code is no different. Many scientists use code to do data analysis. This process begins with the acquisition of data, either by running an experiment or model that generates data or by identifying observational data that may be useful to test a hypothesis. Next, the data is analyzed. It is very likely that the code required to read or analyze a new data set was already created by someone. The existing code might require some degree of modification to meet a researcher's unique parameters. Even the development of a new model can incorporate specific elements of existing code from different sources.

Understanding how to find and use others' code, create your own, and share it is an important part of advancing open science. Just like good data management practices, knowing some of the details about how to share it will not only help you use it later, but also help others understand how to use and cite it so you get credit!

Code example from https://github.com/UCB-stat-159-s23/site/blob/main/lectures/climate-data.ipynb

17.5.1 What is Code vs Software?

When we write "software," we are actually writing text code and using an interpreter or compiler to translate it into a program that the machine can run. Code is a language that humans can type and understand. Software is often a collection of programs, data, and other information that a computer system uses to perform specific tasks. An example is a software library, which is a suite of data and programming code that is used to develop software programs and applications.

Often, scientists write and publish code that helps others reproduce their results rather than creating software packages. But many scientists aren't starting their code from scratch. There are large open- source software libraries that scientists use and contribute to, such as scipy, astropy, matplotlib, and others. These libraries let everyone do science faster and better

because they have been written, tested, and are used by thousands if not hundreds of thousands of people. These libraries have been widely adopted because they are open-source – which makes it easier to collaborate with anyone, anywhere.

17.5.2 What is Open Source Software

Open-source software is distributed with its source code without cost, making it available for others to use, modify, and distribute with its original rights and permissions.

Often, open-source software is transparently shared in a public repository, and sometimes maintained through collaboration. Open-source software development is the basis for a vast range of research software packages.

There are a variety of license choices that can be made for open software which can allow the creator to retain various levels of ownership and rights. The choice of license impacts reuse by others. But first, let's break down the main types of software scientists use based on their purpose by showing examples of each type.

17.5.3 Types of Software

Scientists use and produce a wide variety of different types of software during projects. While many researchers might just use equations in a spreadsheet, others may use open source libraries for advanced machine learning model development and plotting results, while others may contribute to open-source libraries in their field and grow their reputation and impact that way. Here are some examples of different types of software that you might encounter.

General Purpose Software – General purpose software is produced for wide use and not specialized scientific purposes. This includes both commercial software and open-source software. Many widely used productivity software packages are open- source success stories:

- Linux kernel, GNU userspace, and various Linux and UNIX distributions
- PostgreSQL – open source enterprise-grade database
- WordPress and Apache web hosting tools
- Firefox and Chrome
  - Chrome's engine is Chromium which is forked from WebKit which was forked KHTML. This was possible because it had a license that allowed for this type of reuse. All major browsers today except Firefox can be traced back to KHTML.
- Android operating system among others
  - You can look at the Android source code, but you can't modify it and install it on a device. And even if you could, you couldn't use any of the standard services (e.g. Google Store) with that. So it's "open" in the same sense that last night's lottery numbers are "open".

Operational Software – Operational software is used by data centers and large information technology facilities to provide data services. For example:

- Fprime – Space mission flight software

Infrastructure Software – Infrastructure software is used by data centers and large information technology facilities to provide data services. Examples include: - PODAAC – Distributed archiving and processing software - UFS – Operational weather forecasting model software - Metadata Compliance Checker, APIs, Web apps, Giovanni, McIDAS

Libraries – Libraries are generic tools for implementing well-known algorithms, providing statistical analysis, or visualization which are incorporated in other software categories. Examples include: - NumPy – Scientific computing with python - scikit-image – Image processing algorithms in python - deal.II – Library of algorithms to solve partial differential equations with finite elements

Modeling and Simulation Software – Modeling and Simulation Software either implements solutions to mathematical equations given input data and boundary conditions, or infers models from data. They often use libraries. Examples include: first-principles models, data-assimilation tools, empirical models, machine learning, mission planning and engineering tools, among others. - OpenFOAM – Computational fluid dynamics software - MOM6 – General ocean circulation model - ASPECT – Planetary convection software - Atmospheric radiative transfer, stellar evolution, upper ocean turbulence, solar wind predictions, orbit propagation (e.g., OpenGGCM, MESA)

Analysis Software - Analysis software is developed to manipulate measurements or model results to visualize or gain understanding. This software often evolves from single-use utility software and may incorporate libraries. - Photutils – tools for detecting and performing photometry of astronomical sources

Single-Use Utility Software – Single-use utility software is written for use in unique instances, such as making a plot for a paper, or manipulating data in a specific way. This code often uses libraries for analysis, plotting, or reading data. This software is the most common type that gets included into Open Science and Data Management Plans (OSDMP), which we will talk about shortly. Examples include: - Angus et al. 2019 – Fitting a gyro relation to Praesepe - Webb telescope spots CO2 on exoplanet for the first time: what it means for finding alien life. All the data and models presented in this publication can be found here. - Constraining the increased frequency of global precipitation extremes under warming - Code at: https://doi.org/10.5281/zenodo.6288035 (2022)

17.6 Principles, Benefits, and Challenges

17.6.1 Principles of Open Code

Open software principles are derived from open-source software best practices. They establish guidelines that advance open science and aim to enhance the value and impact of research.

Transparency	Whether you are developing software or solving a business problem, we all have access to the information and materials necessary for doing our best work. When these materials are accessible, we can build upon each other's ideas and discoveries. We can make more effective decisions and understand how those decisions affect us.
Collaboration	When we're free to participate, we can enhance each other's work in unanticipated ways. When we can modify what others have shared, we unlock new possibilities. By initiating new projects together, we can solve problems that no one can solve alone. And when we implement open standards, we enable others to contribute in the future.
Share early and often	Rapid prototypes can lead to rapid discoveries. An iterative approach leads to better solutions faster. When you're free to experiment, you can look at problems in new ways and seek answers in new places. You can learn by doing.
Inclusive	Good ideas can come from anywhere, and the best ideas should win. Only by including diverse perspectives in our conversations can we be certain we've identified the best ideas, and good decision-makers continually seek those perspectives. We may not operate by consensus, but successful work determines which projects gather support and effort from the community.

Community	Communities form when different people unite around a common purpose. Shared values guide decision making, and community goals supersede individual interests and agendas.

Credit: The open source way | Opensource.com

Sharing code enhances science because it enables reproducibility, reusability, and replicability. The decision to share code benefits the scientific community because it increases transparency, participation, and collaboration. Sharing code at any point in the research process can be valuable.

In most cases, the source code used to generate results in peer-reviewed papers should be published, cited, and accessible.

17.6.2 Benefits of Moving to Open Software

Science moves faster when researchers are able to work together, help correct errors, build on each other's results, and share resources. Sharing software is a key part of open science that:

- Accelerates science by making it easier to use and build on software developed in previous work.
- Minimizes the time and cost of repeated development of similar software and the reproduction of scientific computations.
- Increases the potential number of users and developers and thus helps improve quality and trust in the software.
- Increases the likelihood that developers gain visibility, sustainability, software quality, and advance their employability.

17.6.3 Challenges of Moving to Open Software

It is not uncommon for research groups to spend years developing code, writing papers with the results, and gaining scientific influence by not sharing the code. Anyone new who wants to work on a similar project is at a huge disadvantage because they would have to start from scratch. Also, anyone wanting to work in that area is forced to collaborate with the group. This group retains a very real competitive advantage by keeping it closed source. However, this approach stifles innovation and hurts scientific progress. Many funding agencies are now requiring that code is shared at the time of publication, if not before. But challenges and fears remain:

- Openness has costs: time spent documenting, publishing, responding to users/maintenance and cleaning up/enhancing quality.
- Effort is required to learn how to leverage the new tools and knowledge (resources are available to ease this effort).

Fear	Discussion/Mitigation:
Scooping: What if someone re-uses my code to publish a result I was working on?	Yes, this can happen. But, in many fields, if it is clear that someone is actively working on a problem, the decision by another scoop may have a short term gain but long-term loss. In the scientific community, reputations serve as a cultural currency and being collaborative generally leads to increased career successes. If you are sharing your code, ensure it has a digital object identifier (DOI) so you get credit. This does not prevent anyone from using it or extending your analysis, but it does ensure you will get credit for your contribution. There is a nice article about this here.
Misinterpretation or misuse	Provide sufficient contextual information (documentation) to allow others to understand your code fully to reduce this risk.
My code will be used, but not cited	While it is not common for researchers to cite code, data, or other non-published articles, science ethics dictates that you should be cited if your work is used. Remember to appropriately cite the material of others so that you're not adding to the problem.
Code is too sensitive to share	User controlled access to help maintain sensitivity and security.
It won't be useful to anyone else	You never know how materials might be used. Individuals who contributed a wide variety of seemingly unrelated software projects ended up helping NASA land a rover on Mars!

17.6.3.1 Ultimately, you are free to deploy the open software principles and resources in your research to maximize its impact and meet the expectations of your sponsors and community while managing costs.

17.6.4 Activity 1.1: Relating Principles to Benefits and Challenges

Determine whether a statement is a benefit or challenge by dragging each to the correct box.

Benefits

Makes it easier to use and build on software developed in previous work.

Users are free to use and modify Open Software minimizing duplicated effort.

Can increase usage of the software, which can help improve software quality.

Open Software developers can gain visibility & sustainability of their software.

Challenges

Requires extra time for activities like documenting, publishing, & maintenance.

Effort is required to learn how to leverage the new tools and knowledge.

Key Takeaways: Relating Principles to Benefits and Challenges

- Making software more open by following the principles has benefits and challenges, which are related.
- Greater benefits typically come with greater challenges.
- In most cases, individual scientists and society will both benefit from more open software.

17.7 When Not to Share

There are valid reasons that restrict a researcher's ability to share their complete code or software packages. Some of these reasons may include:

- The code incorporates a country's military secrets or its dissemination violates national interests or security concerns.
- The code incorporates intellectual property or patented data and information.
- Institutional policies or organizational regulations do not permit the sharing of code.
- Think about what you are sharing and the implications of sharing it (for example - do you have permission from everyone involved?).

17.7.1 Licensing Code

The collaborative data science handbook by The Turing Way says of restrictions to open source sharing, "As with anything else in society, some of what you can and cannot do in software (or hardware) development is determined by the law. Licensing is therefore an important aspect of sharing/publishing open source projects as it provides clarity for anyone looking to reuse an open source project. Without licenses in place, anyone who wants to reuse it will be left with legal ambiguity as to the status of using your intellectual property."

To be considered open source, software requires a license that complies with the Open Source Definition. One criteria of this definition demands that open source licenses "must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software."

In the next lessons, licenses will be discussed in more detail. As you are working on a project, you may want to use code developed by others, develop your own code, and then share it. Licenses affect all aspects of this process and it is important to understand how different licenses may affect your ability to share your code at the time of publication. It is also important to consider any requirements from your funder or institution about how you license your software.

17.7.2 Planning for Openness: Using the Use, Make, Share Framework for Open Code

Funding agencies and journals are increasingly requiring researchers to share software.

For example, NASA's ROSES, which solicits Earth science research proposals, requires researchers to make their software publicly available:

17.7.2.1 "Data and software developed using Research Opportunities in Space and Earth Sciences (ROSES) funding in support of a peer-reviewed publication shall be made publicly available at the time of publication"

https://science.nasa.gov/researchers/sara/faqs/osdmp

Planning for a research project requires researchers to determine their mode of collaboration and method of sharing code. This step is often documented in a Software Management Plan (SMP) within a research proposal. An SMP details the what, when, where, how, and who will be sharing the code or software.

17.8 Software Management Plans (SMP)

Software management plans encompass both code and software.

What?	Description of types, management, preservation, and release of software.
When?	The schedule for software archiving and sharing.
Where?	Location where software will be shared and archived over the long term.
How?	Enable reuse of software through assigning a DOI, license, contribution guidelines, etc.
Who?	Roles and responsibilities of the team members.

As your research starts using, creating, and sharing code, the SMP provides a guidebook for everyone on the project that establishes a common understanding.

Is your project sharing all code publicly or just code that goes into a publication? Will your team be contributing back to open-source projects or just writing code that builds on them to produce results? Considering these questions early will influence how much time and energy you may want to spend on documentation and how you plan to share the code.

17.8.1 Open Code is a Spectrum

Just like data, code can be shared in many different ways to increase reusability. Code can be shared without any documentation, purely as a reproducibility artifact, or code can be well-written, documented, and openly-licensed to maximize re-use. Both of these approaches have value and depend on the time, energy, and funding that researchers have available.

- There is a spectrum of openness when it comes to open software that ranges from open-source software to closed source software.
- An example of something "in between" could be an executable file with documentation on how the code works.
- Some projects may be open from inception and continuously share all code throughout development. Others may share some of the code at the time of publication. Other projects may only make code available once funding ends. A variety of valid reasons factor into a project's approach to sharing.
- While some factors restrict the degree of openness that software can be, each step towards sharing advances the open science movement.

- By sharing more ideas and software, communities have driven creative, scientific, and technological advancement faster than the restricted pace of closed science. Peer production and mass collaboration creates more sustainable software development.

While researchers and institutions may not be able to share all their code, they can make efforts to shift on the openness spectrum from closed code to open-source code and software.

In the activity below, drag each slider to explore the spectrum of openness.

17.8.2 The Practice of 'Open'

Review how the key tasks in the software development life cycle are covered in the "Use, Make, Share" framework flow.

As with open data, different aspects of open software are described in terms of Using, Making, and Sharing of open software.

A key difference with software is that the process is typically more cyclical and repetitive than with data or results. Typically, software constantly evolves. Thus, the boundaries between "Use-Make- Share" are less rigid and the process is typically more dynamic and circular than pre-planned/fixed and sequential.

17.8.3 Activity 1.2: How Can You Use Open Software in Your Work to Advance Open Science

In this activity, you are asked to reflect on how you have used and can use the open software principles to advance your work.

Consider the following questions:

1. Have you used open software principles 1 in your work?
2. What are some of the successes and challenges you have encountered?
3. What resources did you find useful for advancing open software in your work?

17.8.3.1 Key Takeaways: How Can You Use Open Software in Your Work to Advance Open Science

- Open software is a collaborative activity.
- We can all learn and benefit from each other in making our scientific software more open.

17.9 Lesson 1: Summary

In this lesson, you learned that:

- In open-source software, anyone can see the underlying source-code.
- Open-source principles promote transparency, collaboration, sharing, inclusiveness, and communities.
- Open-source software accelerates science, minimizes time and cost of repeated development of similar software and reproducing scientific computations, and can improve quality and trust in science.
- Licenses for open-source software dictate its shareability and reusability to developers and prospective contributors. Funding entities and affiliated institutions may impose restrictions on how developers license their software.
- A software management plan (SMP) is a project guidebook with a common understanding of data management practices that a research team can work from.

17.10 Lesson 1: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/03

Read the statement below and decide whether it's true or false:

Software is referred to as open source when it is publicly accessible; anyone can see, modify, and distribute the code as they see fit.

- True
- False

Question

02/03

Which of the following are valid reasons why scientists keep their source code closed? Select all that apply.

- National security concerns
- Institutional policies
- Data privacy concerns
- Attribution concerns
- Quality concerns

Question

03/03

What are the main sections in a software management plan?

- Types of code and software
- Schedule for sharing software
- Where software will be shared and archived
- What license it will be assigned
- Roles and responsibilities of team members
- All of the above

18 Lesson 2: Using Open Code

18.1 Navigation

18.2 Overview

In this lesson, you learn the steps for using existing open code in your work. These steps include discovering, assessing, reusing, citing, and acknowledging.

18.3 Learning Objectives

After completing this lesson, you should be able to:

- Describe the process of using open code and list some key elements of discovering code.
- Describe the four key considerations when assessing open software: functionality, interoperability, security, and licenses.
- List some common problems that arise when reusing Open Code and best practices to resolve them.
- Describe how, where, and under what circumstances one should acknowledge (cite) code.

18.4 Discovering Open Code and Software

Many people discover code through discussions with their colleagues or by reading journal articles and attending talks at conferences. This is a great way to find out about code that might have applications for your scientific problem.

What other ways can someone search for open code? As a first step, look for code that already exists because chances are that someone else has already had a similar problem and published their code online. A common way to search for existing code is with a general search engine. Search engines offer one indicator of a code's relevancy, how recently it was updated, and how frequently others reference it.

Example	I'm a new graduate student starting to work on modeling turbulence in the Southern Ocean to better understand sea surface temperature (or ocean heat uptake) and climate change. Is there some software available to model how eddies in the ocean affect sea-surface temperature?
Exercise	General Search on the term "Software for ocean turbulence modeling"
Result	General Ocean Turbulence Model (GOTM)

This successful search is predicated on the developers of GOTM making their code open.

18.4.1 Open Software Discovery Depends on Developers Following FAIR Principles

Discovering open software depends on developers making their software easy to find. The Findable, Accessible, Interoperable and Reusable (FAIR) Principles for research software suggest:

- Software and its associated metadata must be easy for humans and machines to find.
- Software must be described with rich, searchable, and indexable metadata.
- Software must be findable from all relevant search points

Reference: "The FAIR Guiding Principles for scientific data management and stewardship" Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). See also Module 1.

However, you may have more specific needs. The following sections cover additional ways to help discover relevant software that meets specific research demands.

18.4.2 How to Search for Open Code

A successful search for open code demands a clearly defined purpose. Developers must first determine the tasks they expect their code to carry out. The requirements associated with these tasks can determine the best suited programming language.

Next, familiarize yourself with the terminology of others who created open software with similar requirements to your own. The keywords affiliated with your programming purpose or requirements can serve as a starting point when searching for relevant code. These keywords can be found in community forums about open source programming and in related scientific journal articles. With adoption of open access principles by many academic journals, prospective programmers can peruse scientific papers from fields related to their research in order to find, and sometimes make use of, existing code that will fulfill their requirements.

18.4.3 Know Where to Search

The open software ecosystem is vast, organic, multifaceted, and highly distributed.

If you are looking for scientific software, community standards increasingly require code to be published and linked to scientific papers.

Thus, the scientific literature and its ancillary code archives are increasingly a great place to look for scientific open code.

Most open code is not developed by or for scientists. However, open code enables research every day.

18.4.4 Where to Look Depends on What You Need

There are several popular search engines for code snippets. First, you can simply search on Google. Other commonly used search engines include GitHub Code Search and Stack Overflow. These search engines allow you to search for specific code snippets by programming language, keyword, or other criteria. GitHub Code Search allows you to search GitHub, a popular code repository for scientific software. Stack Overflow allows you to search forums, where users discuss solutions to coding problems.

18.4.4.1 Examples of code repositories:

GitHub

GitLab

Bitbucket

Example - GitHub Code Search

In this example, we will practice searching for open access code on GitHub. Let's work through a scenario in which you would like to search for the Lomb and Scargle method for estimating a power spectrum.

Example background

GitHub enables users to collaborate on a shared project and track their changes with version control. Users can create a repository and grant others access, or make it open access. GitHub involves a large community of open access users who make their code available for free.

Example instruction

Begin by visiting the GitHub website to search for openly available software packages. You will need to create a free account for this action. Navigate to the Search Code page to begin your search and access tutorials on the interface and capabilities of the search portal. Alternatively, you can simply input your search terms in the search bar while on your profile page. Next, input the related keywords into the search bar. Search for "Lomb Scargle" and find several repositories with relevant code in various languages, along with thousands of related snippets of code. Congratulations! You have begun your open access software journey and can now view the work of thousands of others who once were where you are now. Upwards and onwards!

Screenshot of the repositories returned from our search

Screenshot of the code snippets returned from our search

With open software, knowing where to search and what to search for can be a challenging problem. You can always start with a Google Search. However, it can be valuable to think through some of the questions that guide the discovery process. If the user lacks relevant experience, it can also be helpful to engage experienced colleagues at this stage.

Review the flow chart that illustrates how the search follows the definition of the need.

18.4.5 Open Software is Aggregated and Searchable in Repositories

A software repository is an online collection of stand-alone application software packages. Repositories typically control access and track the deployments/downloads of packages.

Software packages are often provided as executables without code.

The collection typically includes metadata, documentation, and licensing restrictions on each package. It may include different software package versions and the platforms or environments on which the software package can be executed.

Most research code should be open source software, which is stored in code repositories.

18.4.5.1 Examples of software repositories are:

Software Heritage

Open Source Development Network (OSDN)

SourceForge

Free and Open-Source Software Hub (FOSSHUB)

Googlecode

Comprehensive Perl Archive Network

PyPl

CRAN

NASA Resources for Discovering Open Software

These are a few links to NASA-specific repositories that may be of interest: - NASA Open Source Software - NASA Open APIs - Science Discovery Engine A strophysics Data System - Earthdata Developer Portal Exoplanet Modeling and Analysis Center

18.5 Assessing Open Code and Software

So, you've discovered some exciting open code that might help you solve your scientific problem. Can you trust this code you discovered on the web? Will it be useful? How much time will it take to learn it? Could the code contain malware? Could you get in legal trouble for using it?

Examples: You found the "General Ocean Turbulence Model (GOTM)" on the internet, and it looks promising. Or, you just found lots of code snippets and functions related to the Lomb-Scargle power spectrum. Now you would like to assess these pieces of code to help you decide if you should use them. This section discusses some best practices for assessing if the code will help you.

18.5.1 Four General Considerations for Assessing Open Software

Software assessment criteria are similar, for any level of openness:

- **Functionality:** Will it be useful for your scientific problem?
- **Interoperability:** How hard will it be to use?
- **Security:** Is it safe? Would using the software create a security risk?
- **Licenses/restrictions:** Can you use it? Is it legal to use the software in your project?

18.5.2 Functionality: Assessing Scientific Utility

18.5.2.1 Does the software meet your scientific needs?**

- Does it address your specific science question?
- Do studies similar to yours use it?
- What papers cite it and how do they use it?
- Talk to your advisors or colleagues that might have experience with it.

18.5.2.2 Testing the scientific compatibility

- Does the software contain scientific test cases? If so, reproduce a case that is applicable to your problem; make sure the results are as expected.
- If you've done similar scientific analysis/modeling previously, reproduce your prior results with the new software. Are the results consistent?
- Incrementally modify a given test case to address new scientific questions. Alternatively, develop your own case, if necessary, following relevant examples.

18.5.3 Interoperability: Ease of Use

18.5.3.1 Is the code written in a language that you are familiar with?

It can be easier to use coding languages that you are familiar with, then import the code into existing software rather than try to use a new language. On the other hand, the use of existing packages and executables can accelerate your work.

18.5.3.2 Check for good documentation

Read the README file. Does the software meet your functional requirements? Are the environmental dependencies well-defined and reasonable?

18.5.3.3 Check the evidence of interoperability with other projects and codes

It is a good sign if you can find evidence that the code has been used successfully by other users that have similar scientific or technical needs.

18.5.4 Factors for assessing the quality of open source software

To quickly assess the community usage and quality of software repository, use the tools from the repository where you found it. GitHub, for example, permits a quick scan of development activity as evidenced by the number of times the code has been downloaded or 'forked' in GitHub parlance. You can also view the amount of activity in a community. GitHub also provides insights into the quality of the software.

18.5.5 The Importance of the README File

- Example above: Astropy
- Always the starting point when assessing software.
- Explains what the software does, how to install and use it, or points to files with that information.
- Assumes limited prior knowledge by the reader / potential user.
- Includes a compatibility description, e.g., dependencies.
- Includes usage examples and/or test cases.

18.5.6 Security: Considerations When Using Open Code

You have found some Open Code that will help you solve your scientific problem and it looks easy to use. However, you may still have some reservations. Perhaps you are unsure if the code poses a security risk, for example.

The risks are relatively low for small snippets of code that are easy for you to fully understand. However, you may not be able to fully understand all components of a large Open Software Package.

Open software is perceived to have more security risks. This is generally less of a problem for open source code than executables because the code can be audited for security vulnerabilities by the community. How can you assess security in this case?

- Consult with your institutional open software policies and IT staff
- Use authoritative reputable sources to minimize security risks
- Set strict security rules and standards when using a dependency
- Use security tools to check for vulnerabilities (e.g., Open Worldwide Application Security Project®)
- Avoid unsupported open-source software. Switch to actively developed components or develop it yourself
- Check with your latest institutional policies on using Machine Learning and Artificial Intelligence tools
- Use caution when using external tools with secure or closed access data. It may be possible for the external tool to publicly share what should be restricted information

18.5.7 Licenses

So, you want to reuse some open code you discovered. It is essential to check the legal restrictions and requirements imposed on users, which are generally provided in the license.

Although licensing is a nuanced subject that you will learn more about in Lesson 3, it is useful to be aware that there are generally two classes of license: permissive and non-permissive. Permissive licenses, most commonly Apache 2.0, MIT, or BSD, will generally allow you to use the code for your scientific research with little restriction, whereas non-permissive licenses such as copy-left licenses, impose substantial restrictions on how you use the code and require more careful consideration.

18.6 Reusing Open Code

Software can be reused in a variety of ways. A software package can be executed on its own to provide a complete analysis or models depending on the input parameters. Alternatively, the package could be imported as part of a larger library to provide specific functionality. Also, code snippets can be copied into existing code, if permitted, or the code could be re-written and incorporated into new software.

If you simply intend to reuse a code snippet, continuously test that your selected code works as you expect. If you are reusing a more complex code, there are additional considerations.

18.6.1 Selecting the Appropriate Version for Reuse

Consider the following when selecting among multiple versions of open source software.

Use the latest stable release when possible	Just like software updates to your phone or computer's operating system or apps, it is important to use the latest stable release. Developers often release developmental versions that include new features or bug fixes that are not fully tested. For this reason, using a developmental release is generally not recommended.
Determine the origin of the version you intend to use	Determine whether the version you intend to use comes from a modified open-source project or from its original source project. With this information, determine which source is more appropriate for your project.

Check for issues and bugs	Check for any known issues or bugs with your selected version that could cause problems. Find current information on issues or bugs by checking release notes, issue trackers, and developer forums.

18.6.2 Resolve Problems in Reusing Software

- Implement tests to verify that the software performs as expected in your application.
- If you run into problems, revisit the release notes, issue tracker, and/or user/developer forums.
- Don't be afraid to ask experienced colleagues for help.
- It is better to seek and obtain help in a public forum than in private (eg. email). Part of open science is working in the open. Often you may find through a search that other users have similar questions. Someone may have already offered a solution. If not, it is likely that others will benefit from your question being answered in public.

18.6.3 Activity 2.1: Ways to Get Help Using Open Software

In this activity, you are asked to select from a list of ways you can resolve some common problems that arise when using open software.

18.6.3.1 Exercise 1

Select how you can resolve this problem when using open software: Difficulty finding open software that meets your needs.

Select all that apply. - Reach out to expert colleagues - Read related peer reviewed literature - Conduct a search of various popular repositories - Read the README file - Read the license file

18.6.3.2 Exercise 2

Select how you can resolve the problem when using open software: installation difficulties.

Select all that apply. - Reach out to the developers on a public forum - Read related peer reviewed literature - Conduct a search of various popular repositories - Read the README file - Read the license file

18.6.3.3 Exercise 3

Select how you can resolve the problem when using open software: software is not working as expected.

Select all that apply. - Reach out to the developers on a public forum - Read related peer reviewed literature - Conduct a search of various popular repositories - Read the README file - Read the license file - Consult the release notes, issue trackers, and public forums

18.6.3.4 Exercise 4

After answering the questions above, work through some specific examples of how you would resolve problems on your own. For example, navigate to the astropy code repository on GitHub or another repository of your choice, and find the README and LICENSE files. Determine how you would contact the developers for help, etc.

18.7 Citing and Acknowledging Open Code Use

Imagine that you've used Open Code pulled from the web and it made a big difference for your project research paper. How should you provide due credit for the open access code that contributed to your research?

Example: You managed to implement GOTM to learn something new about ocean turbulence in the Southern Ocean, or you managed to compute a Lomb-Scargle periodogram using astropy. Here are some questions to consider:

18.7.1 Should you cite the Open Code?

Cite any code that you view as having contributed to your research:

- Did the code play a critical part in your research?
- Did the code provide something novel?

In most cases, a code snippet on Stack Overflow does not constitute a citable research contribution. However, an author can still decide to cite it if they chose.

Instances when shared code directly impacts the scientific results and requires a detailed description include:

- Numerical modeling or simulation
- Automated analysis, such as image processing or optical recognition

See the journal where you are publishing if they have any specific instructions on how to cite software (e.g., AAS Software Citation Suggestions).

In some cases, a software's licensing terms and conditions require acknowledgement or citation in the references or bibliography of any publications based on research that made use of the software.

18.7.2 How to cite?

Ideally, use and cite code that is archived in a long-term repository with a persistent DOI. Follow the guidance about the preferred citation format, which is provided in the long- term repository and may appear in a README or a CITATION file.

DOIs provide a persistent identifier/link for research outputs. Thus, it is preferable to cite code in long-term repositories linked to a DOI. URLs (e.g., Stack Overflow) and active repositories (e.g., on GitHub) are mutable but can be used if there is no alternative.

Packages may provide a way to cite individual versions as well. For reproducibility, cite both the overall package and the version that is used in your work. As functionality of a package may evolve with the release of new versions, this helps provide a specific description of your work.

If you are writing software, you can also cite in the comments and documentation of the software that you have used.

18.8 Lesson 2: Summary

In this lesson, you learned that:

- Open code exists in a vast, organic, and distributed ecosystem. Discovering Open Code depends on defining your requirements, knowing where to look, and developers using FAIR principles.
- Scientific papers are now a good place to discover scientific Open Code, since many journals require the code used in the paper to be linked via a DOI.
- Before use, it is important to assess open software for functionality, quality, interoperability, security, and license/reuse restrictions. Your first step should be to look for a README file.
- When reusing open software, use the latest supported version and test the software to ensure it functions as expected. If problems arise, reach out to the developers or user community, ideally via a public forum.
- It is important to cite and acknowledge open software that significantly contributes to your work, as well as share your lessons learned and any contributions with the developers and user community.

18.9 Lesson 2: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/03

Discovering open software successfully depends on which of the following:

Select all that apply.

- Well defined requirements
- Knowing where to search
- FAIR open software exists to meet your needs
- All of the above

Question

02/03

Read the statement and decide whether it's true or false:

It is best to reach out to the developers of open access software via private communication if you run into problems.

- True
- False

Question

03/03

When citing Open Code, it is best practice to cite:

- The primary working repository, e.g. on GitHub. It has the most recent version of the code, including any updates since your paper was written.
- A long-term code repository linked to a DOI, e.g. on Zenodo. This repository contains a persistent version of the code that you used.

19 Lesson 3: Making Open Code

19.1 Navigation

19.2 Overview

In this lesson, you will learn about the practical steps to make code openly accessible. Large volume and well-established software have different needs than an incipient project. For example, a script written to create a simple plot has different requirements than a software package that models the Earth's climate. The size of a research team can also determine the steps required to make code open access. This lesson covers the process to make code usable to other researchers through documentation, considerations around licenses, and software development best practices.

19.3 Learning Objectives

After completing this lesson, you should be able to:

- Describe the key considerations when planning a new open software project.
- List three reasons for projects to use version control.
- Explain the purpose and recall general information typically included in a README file.
- Be able to select a license for your code and list the differences between permissive and protective open-source software licenses.

- Explain best practices in software development that support transparency, inclusion, and reproducibility.

19.4 How do We Plan for Making Code?

Code is written to solve a challenge. This can range from producing a plot, to data processing Earth observations, to modeling the Universe. The challenges associated with writing code can range in difficulty, from simpler tasks such as the use of spreadsheets to more complex activities such as the creation of extensive libraries and use of high performance or cloud computing. Code can be developed as an individual, team, or a community. Once written, code might be used for decades or never again.

When starting a research project, it is useful to answer the following questions:

1. What problem am I trying to solve, and are others in my community facing it as well?
2. Are there existing solutions? (In Lesson 2, we explored how to look for existing solutions.)
3. Did you find code that was close to what you want but didn't quite meet your needs?

You could potentially contribute to it instead of writing something new.

Even if a solution already exists, there might be good reasons to develop your own code. Instances include:

- The code is written in a different programming language than you are familiar with.
- The license is not open enough to adopt it.
- To try new techniques or to develop a deeper understanding of the problem.

It might take more time to start a new project, or it might take more time to integrate someone else's code than writing your own. You will have to make that call.

We looked for existing code, and though we found a few things that were close we decided in the end our needs were unique enough - we're starting a new project!

19.4.1 Starting a New Project

When starting a new project, the key things to consider are:

1. Define the project scope, its primary features and any limitations, and the intended audience.
2. Consider resources required for the software to run. Will it be on a personal computer, a high-performance computing server, or on the cloud?
3. How will it be managed?

This lesson focuses predominantly on the question of how to manage open access code.

Who will be working on the project? What are some of the development best practices? How will you share it openly? How will it be licensed?

19.4.2 Organizing a Project

Source: https://xkcd.com/1459/

Software projects can be organized in a variety ways, each that involve unique considerations about how to begin. Many projects start out as a single script that was only intended for a single use. However, a script can grow into a much larger project with unforeseen applications in its original or new field of research. Other projects can start with formal requirements and standards.

Making code public has many advantages:

- It enables open collaboration.
- It invites constructive feedback that contributes to a code's accuracy and robustness.
- People with less experience with the subject matter will learn more.
- Those with less programming experience can learn from those with more programming experience as they improve the code.
- It provides an intermediate product that can still be cited.

When naming a project, conduct a quick search of the envisioned name to see what shows up. Avoid names with many other uses as this will make it difficult for others to discover the code. Also, do not choose embarrassing or trademarked names.

Hosting the product on a version control platform ensures the permanence of your project. If code only exists on your computer, it may disappear if the computer is damaged or is lost.

Documenting the production and management of your code benefits both you and those that might use your code in the future. You are your own best collaborator. Documentation can save you from a headache should you reuse the code in six months or attempt to recall meticulous details about your process later on.

Questions to consider when choosing a programming language:

- Will potential collaborators be able to contribute in the chosen language?
- Which languages are you most experienced with?
- Are there any limitations from your computing environment that would impede your ability to write or manage this code?
- Languages have strengths and weaknesses; which are most important for your project?

Before someone else can use your code, they're going to ask some questions:

- Where can I find your code?
- Is your code documented?
- In what ways am I allowed to use your code?
- Will you accept changes to your code? If I find a bug, what do I do?
- How do I trust your code works?
- How do I know if the code will be supported long term?

19.5 Importance of Version Control

Your code will change significantly over the lifetime of your project. Just as we appreciate the ability to track earlier versions of documents or versions created by different people, inevitably someone will want to be able to revert, compare, and synthesize changes in code.

The most popular tool for version control is git. Git is a system that tracks changes in computer files, similar to Google Docs or SharePoint but more applicable to code script. Git is usually used in conjunction with a version control platform such as GitHub, Gitlab, or Bitbucket. These tools were covered in Module 2.2.

Version control enables the following:

- Helps developers keep track of changes to a project's code (as well as supplemental files and documentation) over the entire course of a project's evolution.
- Revisions to a project's files can be tracked, including contributions made by different people.
- Undesirable changes (like errors or bugs) can be reverted at any time.

Version control is a good practice for coding, even if you are not immediately sharing the code. You can use version control with your code privately on your computer, or use the private mode on hosting services (e.g., GitHub and GitLab). By setting up version control early on, you prepare your code for intended and unforeseen future use.

Further Resources on version control

- Software Carpentry Version Control with Git
- The Turing Way, Version Control
- Use a publicly accessible repository with version control: guidance for FAIR software

19.6 Describing Our Code to Others

19.6.1 README

The first stop for a user when they approach a new project should be the README file. Aptly named, this file contains orientation information that will help a user understand a project's purpose, provides examples of how it can be used, and lists other important information that the creator deems pertinent.

At the minimum, a README should contain the name of the project and a very short paragraph of what the software is. Two to three sentences in a plain-language style that does not assume who is reading it. It's the elevator pitch for the project.

Bad README example	"This code recomputes the fundamental permutation factor of the downward flow (for J < 10, obviously)."
Good README example	"LeapKitten. This Python software package takes any picture of a kitten (JPEG, PNG) and uses artificial intelligence to output what it would look like leaping into the air. In addition, the code takes leap years into account on the timestamp on the image."

In addition, the following information is helpful to add to the README especially if they are not listed elsewhere:

- A list of any code dependencies the software has, e.g. "Numpy, kitten-rng, and human-readable must be installed to run this software."
- How to install and a brief description of how to run the software.
- Detailed description of the software, especially if there is no external documentation.
- Examples of how to use the software.
- Acknowledgement of team members or sources of support.

As seen in these examples, README files can be useful for a collection of scripts supporting a publication or an extensively developed software package.

19.6.2 Contributor Guidelines

The *CONTRIBUTING.md* file gives information about how to contribute to the project. It details how the contribution process works and what type of contributions are needed. While not every project has a *CONTRIBUTING.md* file, the existence of one is a clear indicator that contributions are welcomed.

You'll need to decide for yourself when your project has progressed enough to consider inviting contributors. When it has, create a document called CONTRIBUTING at the top level of your report.

The Astropy contributing guidelines and Numpy contributing guidelines provide two examples.

Bonus Tip: Even if you are developing your code publicly, this does not mean you have to accept contributions from others or maintain your code forever. The contributing guidelines or README are good places to indicate what your expectations are for your code. This can clarify that the code is not maintained or not accepting contributions.

19.6.3 Code of Conduct

The code of conduct sets ground rules for participants' behavior and helps to facilitate a friendly, welcoming environment. While not every project has a CODE_OF_CONDUCT file, its presence signals that this is a welcoming project to contribute to.

19.6.4 Code Documentation

Code Level Documentation for the Developer

Your software should be documented within the source code. Each function should have comments at the start that briefly state, in plain language, what the function is for. This is not only for other developers, but yourself a week later when you forgot what you wrote.

> **Example**
>
> # This function takes the image array and crops it from the center to 50% of the original size.
>
> Without going into details of the data type, calling parameters, etc. this description immediately puts someone looking at the code into the context of what the function aims to accomplish; they can then explore the details.
>
> While you should consider placing a description at the start of a function, use your discretion on where you put similar descriptions of code. At the start of a complex loop or analysis would be good ideas. Don't go overboard - things like this aren't useful:
>
> # set x to 17
>
> x = 17

Descriptive variable, class, and function names can make your code very readable. . Sometimes even great coders are working fast and will name variables 'a', 'temp', or other names that probably won't make a lot of sense in a week or two when they come back to something they were working on. Names like 'baking_time' or 'velocity' are more clear. Variable names should be easy to understand and clearly represent what they are.

Ideally, someone who doesn't write in the software language of the code can read the comments in the file and have a rough idea of what is happening.

Use the comments to put URLs that reference where you might have found the algorithm you're using (e.g. Stack Overflow) or the journal paper where you found the formula you're implementing.

19.6.5 Code Level Documentation for the User

If you are developing code that you expect others to use, produce a manual on how to use the code. As code constantly develops, it is much easier to document while or even before you write any code.

If you write your documentation within the code itself, there are pieces of software that can then extract it, format it, and present it as a polished manual. Examples of documentation generated from the code can be seen for Astropy or NumPy.

They look fancy, but very similar too. These sites were completely generated from comments and documents written in the source code. Different from the comments written for developers of the code above, these comments were written specifically for the audience of external users of the code: the manual.

While there are multiple software packages for automatic documentation generation, the most commonly used ones are Sphinx for Python and Doxygen for most everything else. Markdown is also a popular choice for the formatting language for documentation.

19.6.6 Programming and Documenting

Establishing a Development Environment - Establishing an appropriate development environment will help you write good, clean code and will help you maintain the project as it evolves.

- Configure any necessary tools for writing the code. Perhaps an IDE (Integrated Development Environment) or text editor. Some popular examples include VS Code, Pycharm, R Studio, Xcode.
- Set up a package manager. For example, for Python, one could use 'anaconda' or 'poetry'.

- Create a virtual environment specific to your project to isolate its dependencies (and their versions) from those used for other projects

Structuring Files and Folders - How you structure the files in your project from the beginning will contribute to the success of the final results.

Different programming languages have different standard folder structures. Familiarize yourself with the standards before starting as it will help others collaborate and will likely save you from difficulties later.

There are a variety of sample code structures that can be used to get started. For example, for Python there is Cookiecutter and an Astropy package template.

19.7 What License Should We Choose for Our Code?

19.7.1 Licensing Considerations when Using Open Software

Open-source software licenses are the basis for how scientists use, make, and share code and software. Understanding some of the nuances of these licenses is important because it will affect how your project can license and share code.

A software license is a legal document that states the rights of the developer and user of a piece of software.

An open source license is a type of software license, approved by the Open Source Initiative (OSI) as compliant with the Open Source Definition. An open source license grants permissions for anyone to inspect, use, modify, and distribute the software's source code for any purpose.

Licenses ensure that developers receive credit and control over how their work is used. Without a license, software is assumed copyrighted and without permissions. Programmers include licenses to allow reuse.

Licenses take various forms in order to outline:

- Contractual obligations (if any exist) between the developer and user.
- What the user may do with the software.
- To whom the user may distribute the software (if any such right exists).
- Length of time the user has the right to use the software.

19.7.2 Some Common Types of Software License

Click '+' to travel more information.

Public Domain

Anyone free to use.

Lesser General Domain

Can link to open source libraries, and code can be licensed under any license type.

Permissive

Gives users wide but not complete latitude to reuse/relicense.

Non-permissive

Allows users to reuse, but also gives users the responsibility to share their changes with the community.

Copyleft

Can be distributed or modified if all the code involved is licensed under the same license.

Proprietary

Cannot be copied, modified, or distributed.

Before you choose a license, first check with your organization or employer. They may have specific guidelines about what software license you are allowed to use. Your research grant may also stipulate permissible license types. The software management plan should specify what license you plan to use.

If a license is not shared with a code, a creative work is assumed to be copyrighted by default in the United States. It does not need to be registered, and it is assumed to be automatically protected by copyright the moment it is created.

For software, the license is shared in a file called LICENSE at the top of the repository. It's a standard location people will know to look at. It's not bad practice to put a one line version of the license at the top of each file of code as well, with a pointer to where one could find the full license.

19.7.3 Types of Open-Source Software Licenses

There are two main types of open-source licenses. Permissive and protective (sometimes referred to as copy-left). The difference in these types of licenses is primarily related to the type of license users of the code are allowed to apply to their derivative works.

PERMISSIVE LICENSE

PROTECTIVE LICENSE

The Open Source Initiative defines a permissive software license as a license that guarantees the freedoms to use, modify, redistribute, and create derivative works. An example of this type of license is the Apache 2.0 license by the Apache Software Foundation. It is the most popular and widely used permissive license.

Users have wide latitude for reuse under this license. They are generally free to incorporate the code into their project or use it how they wish. A user of permissive-license open source in a product could redeploy the open source software with a wide range of licenses, including proprietary closed source software.

PERMISSIVE LICENSE

PROTECTIVE LICENSE

Protective (copyleft) licenses are a legal technique of granting certain freedoms over copies of copyrighted works with the requirement that the same rights be preserved in derivative works. This allows users to reuse, but also requires users to share their changes with the community using the same license. An example of a protective license is the General Public License (GPL) that ensures users have the freedom and responsibility to share their changes with the community. It is the most widely used protective license. These types of licenses can result in less re-use by users who may prefer or be required to only use permissive licenses.

19.7.4 Common Licenses for Open Software

Some of the most popular licenses used in open software are:

PERMISSIVE (CAN APPLY ANY LICENSE TO DERIVATIVE WORKS)

PROTECTIVE/ COPYLEFT (ALL DERIVATIVE WORKS MUST DISTRIBUTE ALL ITS SOURCE CODE UNDER THE SAME LICENSE)

Apache License

MIT license

BSD License

PERMISSIVE (CAN APPLY ANY LICENSE TO DERIVATIVE WORKS)

PROTECTIVE/ COPYLEFT (ALL DERIVATIVE WORKS MUST DISTRIBUTE ALL ITS SOURCE CODE UNDER THE SAME LICENSE)

GNU General Public License (GPL)

Mozilla Public License

Common Development and Distribution License (CDDL)

For more information on different types of licenses please refer to the Open Source Initiative OSI.

19.7.5 Activity 3.1: Licenses

In this activity, you are asked to answer whether the following statements are true or false.

Statement 1:

A software license states the rights of the developer and user for a piece of software.

- True
- False

Statement 2:

Without a license, software is assumed copyrighted and without permissions.

- True
- False

Statement 3:

Anyone is free to use software with a "permissive" license without restriction.

- True
- False

Statement 4:

Users are not allowed to copy and modify any software with a copyleft license.

- True
- False

19.8 Programming Best Practices

In this section, some best practices in development are provided including on code review, testing, security, and accessibility. These best practices will improve the quality of code, reproducibility of results, and security of a project. Combined, these actions help improve the robustness of open access code and help to meet the unique challenges that can arise with multiple contributors and revisions that occur over an extended period of time.

19.8.1 Code Review

Code benefits from peer review in the same way as science. Having someone else read over your code and test it is one of the best ways to improve the quality of the code.

Many version control platforms have built in tools that enable developers to review, comment, and iterate on each other's code. These can be done in the open and allow anyone to comment.

Here is a great example of the discussion that can happen when the original creator of an algorithm comments on a python implementation made by a first time contributor to the Astropy project. The open and constructive discussion led to a better implementation of the algorithm along with possible future improvements.

Software packages can be reviewed as their own products as well. Many scientific publications now accept papers focused on software. There are entities like PyOpenSci and the Journal of Open Source Software that provide open peer review of scientific packages. See more details about JOSS in the next lesson on sharing your code.

19.8.2 Testing

A proven method to evaluate the reproducibility of your software is through testing. There are many types of testing that range from testing the smallest testable parts of a code to verifying if a code works as whole under different scenarios. Code testing can include a wide range of different techniques. The following lesson section provides only a brief introduction to the topic.

The main objective of code testing is to evaluate if a code does what its authors intended it to do. Comprehensively testing code can be very difficult as it involves testing the code for generating expected outputs as well as for failing when it should.

SCIENTIFIC VALIDATION

REPRODUCI-BILITY TESTING

BUILT IN TESTS

AUTOMATED TESTING

Whether producing a script or an entire data processing pipeline, the validation of software is critical to ensuring the quality and trustworthiness of the scientific results. This could mean manually calculating the results to check the output of the code or comparing to previously produced results or having another team member test it.

SCIENTIFIC VALIDATION

REPRODUCI-BILITY TESTING

BUILT IN TESTS

AUTOMATED TESTING

Given the same inputs and parameters, can the same results be produced? Making the configuration files, input data, etc. openly available so users can easily run and produce the same published results is a critical way to increase trust in your code.

SCIENTIFIC VALIDATION

REPRODUCI-BILITY TESTING

BUILT IN TESTS

AUTOMATED TESTING

Unit tests enable software developers to bolster their confidence in their code's ability to perform as expected. Unit tests are small functions that sit outside the code base that test a specific function or run a specific test. For example, if a function takes an image and flips it horizontally, one test might check that the resulting image is the same size. Another compares the output using a known image with the expected result. Another checks that a new image is returned.

SCIENTIFIC VALIDATION

REPRODUCI-BILITY TESTING

BUILT IN TESTS

AUTOMATED TESTING

Built in tests can usually be run both manually and automatically. Most version control platforms offer services for running tests automatically. When run this way, code can be checked to see if changes raise any problems. This process of checking the code automatically as it is developed is called continuous development or continuous integration (CI/CD). If a small change made in one part of the code results in an unexpected change in another part, running the tests will uncover this immediately.

19.8.3 Minimizing the Risk of Security Vulnerabilities

Whether using open source, closed source, or commercial software, it is important to consider the security risks inherent in the development of software.

- Ensure minimal, DRY (Don't repeat yourself) code (easier to maintain and fix).
- Use global variables or key managers for credentials. Never include credentials in your code.
- Use well-tested and maintained dependencies. In packages that you maintain, keep the list of dependencies up to date.
- Create software with tools that provide automated scanning and auditing.
- If there are unsupported dependencies that you rely on, assess them to determine how they might introduce security risks and whether it would be appropriate to switch to a different package.

SECURITY TOOLS AND SECURITY VULNERABILITIES

TEST COMPONENTS AND DEPENDENCIES

Commercial and open-source tools have been developed to address the challenge of identifying the security vulnerabilities in different source components. If you do not have any technology to secure your open source usage, you can consider using the Dependabot or OWASP dependency check tools.

The Open Web Application Security Project (OWASP), is an online community that produces free tools and technologies in the field of web application security. OWASP dependency check is a utility created for developers, which identifies project dependencies and checks if they contain any known, publicly disclosed, open-source vulnerabilities.

SECURITY TOOLS AND SECURITY VULNERABILITIES

TEST COMPONENTS AND DEPENDENCIES

Testing the security of the open-source components you are using is the best way to ensure the safety of your applications and your organization. Your commitment to timely and frequent analysis of open-source components should be the same as to your proprietary code.

This is especially true as the component in question may have unknown security vulnerabilities or dependencies that differ with each use case. It is possible for a component to be secure in a particular application but vulnerable in another.

19.8.4 Creating FAIR Software

FINDABLE

ACCESSIBLE

INTER-OPERABLE

REUSABLE

Software includes a persistent and unique identifier and rich metadata, so it is easy for humans and machines to find.

FINDABLE

ACCESSIBLE

INTER-OPERABLE

REUSABLE

Software is retrievable from its identifier via standard communication protocols.

FINDABLE

ACCESSIBLE

INTER-OPERABLE

REUSABLE

Software interoperates with other software; it exchanges data and/or metadata via community standards.

FINDABLE

ACCESSIBLE

INTER-OPERABLE

REUSABLE

Fully described metadata with provenance, meeting community standards. License permits reuse.

19.8.5 Additional Helpful Tips

Here are some further suggestions on how to make your code more accessible, reproducible, and transparent:

Descriptive Names	Variables, functions, and similar entities should be given descriptive names as opposed to vague names. Descriptive names instantly give other programmers an idea of what the variable or function is. For example, the variable name **colourOfCat** is a good name because it describes what it intends to do, which is to encompass the color of a cat.
Metadata File	Consider including a metadata file for your software to make it more discoverable. A 'codemeta.json' can be created using Code Meta's generator to include with your package.
Operation Documentation	Share details about how you are running the code. For example, document the version of a software library you are using, or the version of the compiler. These are often shared in an 'environment.yml' file.
Automation	Consider the following scenario:You are getting ready to publish your paper that includes 17 plots that all depend on a data set released by a mission. Right before you are about to submit, the mission releases an updated version of the data set. How easy will it be to recreate those plots? Software allows you to automate the running of scripts and alert programmers when written so that input files are not hardcoding. This allows programmers to easily re-run code if an initial parameter changes.
Using Standards	Most languages have their own coding style adopted by their respective communities. Following those conventions makes it easier for others to contribute to your code and makes your project more inclusive.
Portability	Share details about how you are running the code, for example the version of a software library you are using, or the version of the compiler. These are often shared in an 'environment.yml' file.

Naming	Many historical terms used in software have negative connotations depending on the context. When considering different terms or naming, consider how different audiences may react to those terms.

19.9 Lesson 3: Summary

In this lesson, you learned:

- Planning a new project requires programmers to have a clearly defined purpose, recognize any resource limitations, and envision a data management plan.
- Using a repository with version control allows developers to track changes across time and from multiple contributors, which can help with troubleshooting for errors and with managing a team of programers.
- A README file should include the name of a project and short but clear description of the software.
- Licenses ensure that developers receive credit and control over how their work is used. Without a license, software is assumed copyrighted and without permissions
- Testing, labeling, and implementing security measures are examples of programming best practices that support Open Science.

In addition to learning how to Share your Code in the next lesson, you will also have some opportunities to put this lesson into practice.

19.10 Lesson 3: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/05

Which of the following should be considered when planning an open software project? Select all that apply.

- The intended user audience.
- What protocol will be used to sync changes between individual contributors and the central repository.
- The programming language to be used.
- Who will financially benefit from sales of the software.

Question

02/05

Which of the following is a benefit of using a version control system in your software?

- New changes are automatically tracked.
- Different contributors can add or edit code at the same time.
- Undesirable changes can be quickly reverted.
- All of the above.

Question

03/05

Select two items that are good to include in a README file from the list below:

- Installation/compilation instructions
- Code development history
- The most important portions of the code
- Usage instructions and example output

Question

04/05

Which of the following licenses allows users to reuse, but also require users to share their changes with the community using the same license?

- Public Domain
- Lesser general domain
- Permissive
- Protective License
- Commercial

Question

05/05

Which of the following practices makes your project more inclusive?

- Including a Code of Conduct.
- Referencing historical events in the name of your project.
- Following standards for the programming language being used.
- Developing the project privately.
- Including a Guideline for Contributors.

20 Lesson 4: Sharing Open Code

20.1 Navigation

20.2 Overview

In this lesson you learn the steps for sharing the software that you developed. These steps include determining if, when, and where software should be shared, which roles are needed, and how to enable others to use the code.

20.3 Learning Objectives

After completing this lesson, you should be able to:

- Describe what it means to share code: for archiving or for code development.
- Evaluate whether you should share your code and list important security considerations.
- Describe best practices for when and where to share code.
- Recall commonly used practices to help others reuse your code.
- List the roles and responsibilities for sharing and maintaining shared code.

20.4 Planning to Share Your Code

"I've been working on code, and now a new collaborator wants to use the code. Awesome! What is the best way to share the code? By email? When should I share the code, and what should I include to ensure the colleague can easily use it?"

20.4.1 What Does it Mean to "Share" Your Code?

There are two major categories of sharing: sharing for development and providing a long-term record.

20.4.2 Open Source Code Development

Writing scientific code is often a dynamic and collaborative process in which multiple people contribute and the code evolves over time. In such projects, it is beneficial to develop open code within a public repository hosting platform such as Github, Bitbucket, GitLab etc. from the beginning of a project. This ensures that all updates are shared openly on the web and can reach potentially interested collaborators and users in near real time.

20.4.3 Archiving Open Code

Archiving ensures your scientific code is accessible for the long-term, and may satisfy archiving requirements from funding agencies and organizations. Long-term accessibility helps others to reproduce your results long after publication. Archiving alone does not promote continued development or collaboration. Archiving is a static and long-term preservation of your software, not an evolution of it.

20.4.4 Should You Share Your Software?

There are several legal and security concerns to keep in mind when creating or using open software.

- Any software you create is usually considered intellectual property and might be controlled by your organization's policies.
- Such policies may influence how openly the software can be shared, and therefore, its license.
- Downloading and contributing to open software projects can be regulated by your organization's IT security policies.

In contrast, if the software was created with external (government) funding, some funding agencies may require the software be openly shared.

20.4.5 Deep dive: Software Management Plans (SMP)

Remember the parts of the Software Management Plan? What do we need to consider when it comes to sharing?

- **What:** Description of management, preservation, and release of software.
- **When:** The schedule for code archiving and sharing.
- **Where:** Location where software will be shared and archived over the long-term.
- **How:** Enable reuse of software through assigning a DOI, license, contribution guidelines, etc.
- **Who:** Roles and responsibilities of the team members.

20.5 Legal and Security Concerns

LEGAL CONCERNS

SECURITY CONCERNS

Anyone writing research code and software should familiarize themselves with their organization's policies on sharing and publishing software. Funding agencies, government or private, may have strict software openness requirements. In other cases, sharing software may not be allowed by the organization.

Legal concerns can include questions such as:

Does a developer or institution own the software?

Does sharing (or not sharing) the software violate the funding agency's policies?

Are there any local laws or regulations in your area that govern the sharing of intellectual property?

What software license is required?

Once you decide to participate in or begin a new open software project, familiarize yourself with your organization's policies and practices.

Find out more about the legal concerns here.

LEGAL CONCERNS

SECURITY CONCERNS

Security is a concern when sharing software. Bad actors can attach malicious code to software in an attempt to infiltrate computer systems through security vulnerabilities, potentially exposing sensitive and proprietary information that can lead to great financial loss for users. Security risks must be considered when sharing software.

Security concerns can include:

Does your organization's Information Technology (IT) policy allow you to checkout the code you want to use on your machine?

Is the repository you want to contribute to reputable?

Are there any open security-related issues with the code?

Once you decide to participate in or begin a new open software project, familiarize yourself with your organization's IT policies.

Find out more about the security concerns here.

20.5.1 Sharing Software Created with US Agency Funding

Many federal agencies are now allowing (if not requiring) the sharing of code created under their grant programs. For example:

- NASA "...we are actively reaching out to projects within NASA to make use of ...resources for publishing open source."
- US Department of Commerce "...requires agencies to develop plans to release at least 20 percent of new custom-developed source code as Open Source Software (OSS) when commissioning new custom software."
- USGS "...software releases are considered to be public domain assets and are generally made available free of restrictions."

Are you funded by a grant? Read the original grant call to see if publishing your code is allowed/required and check whether it has any language about software management and any conditions to publish your code. When in doubt, contact your organization for additional information.

20.5.2 Activity 4.1: Find Your Organization's Software Release Policies

Assume you want to start a new open-source project:

- Find your organization's policies on software releases.
- What is the process for releasing your software?
- Does anybody in your organization have to approve this release?
- Are there any policies regarding external contributors?

- Does your organization require a specific attribution or credit?

20.5.2.1 Key Takeaways: Find Your Organization's Software Release Policies

Software release policies differ by organization and each piece of software is different. Therefore, it is important that we do not make assumptions about the software release policies based on previous experience.

20.6 When: The Schedule for Code Archiving and Sharing

Planning to share your code at the beginning of your project makes sharing easier to do when you are ready. Exactly when in your workflow you decide to publicly share your code depends on your work and the requirements of the funding agency, organization, or publisher.

As an example, what does NASA say?

If you are writing scientific software for a project funded by the NASA Science Mission Directorate then:

"Scientific software needed to validate the scientific conclusions of peer-reviewed manuscripts resulting from SMD-funded scientific activities shall become publicly available no later than the publication date of the corresponding peer-reviewed article. This includes software required to derive the findings communicated in figures, maps, and tables, as well as scientifically useful software from models and simulations."

- **Open-Source Science Guidance**

Other organizations may have different guidance, so it is always best to check what the funding agency or organization requires.

20.7 Where: Where To Share Open Code

20.7.1 General Considerations

Like data, code can be shared in many ways, for example over email or on a personal website, but these methods are not recommended. So, where should you share your Open Code?

First, consider your institutional or funding agency policies that may dictate where you must share and where you can share. For example, some funding agencies specify long-term repositories where your code must be archived, and they may restrict you from sharing in other forms of repositories. Your scientific discipline may have a specific repository for open code.

20.7.1.1 What are some good options and best practices for archiving your code?

- Archive open code with an open access journal article.
- If the open code is in an active online development repository such as Github, then create a version and archive the code at a long-term repository with a DOI such as Zenodo, which can be integrated with Github (more details on this process later).
- Archive the code in other long-term public repositories, such as Software Heritage.

20.7.1.2 Is your code a substantial software package and of interest to a significant number of users from various disciplines? Where else can your open code be shared?

- Develop your software on a public repository such as GitHub.
- Publish to a software repository used by common package managers to make the software easy for users to install (ex. Anaconda, CRAN, PyPI).
- Present the software at conferences.
- Publish the software in a Journal dedicated to open software (ex. JOSS).
- Get your software peer reviewed through communities like PyOpenSci.

20.7.1.3 To share my code, I can just add it to github, right?

Not necessarily. Sharing on a repository is encouraged, but a researcher's funding organization may require a DOI from an archival repository, such as Zenodo, for long-term preservation of your code at the time of publication or version releases.

20.8 How: How to Enable Reuse of Code

Now that you have shared your code in the appropriate way, it's important to consider if you've made it easy for others (or your future self) to reuse your code.

20.8.1 Assigning a License

As you may recall from the previous lesson, assigning an appropriate license is necessary for others to know how to use your code.

As an example, here's how you'd assign a license to a GitHub repository:

Choose the appropriate software sharing license that meets your organization requirements. To create a license template in GitHub, add a new file and type "LICENSE" in the name field, then the "Choose a license template" option will appear.

Make sure that your GitHub repository is public, making it searchable by anyone.

20.8.2 Making the Code Citable

Not all code needs to be citable. When released on its own however, there are a few best practices for how to make your code citable.

Adding code to a GitHub repository is not sufficient for archiving code. To archive, we must assign a persistent identifier.

Producing a persistent identifier for your code is the best way to make it citable. This could take form through a peer reviewed publication that describes the software or by archiving the software with a long term repository that produces a DOI or similar identifier. For code shared on GitHub, a DOI can be easily produced for each release of the software from Zenodo.

20.8.3 Activity 4.2: Create a DOI for a Test Code File

You can create Digital Object Identifiers (DOIs) for your code that makes it citable. You do this by archiving a GitHub code repository at Zenodo and issue a DOI for the record.

Steps for this activity:

Part 1: Create a test public GitHub repository.

1. Navigate to the login page for GitHub and login. If you haven't already, create a free user account.
2. Create a new repository with this link.
3. Type a short, memorable name for your repository. For example, "os-test".
4. Set the repository visibility 'Public' by selecting this option below the repository description.
5. In the following section 'Initialize this repository with:' select 'Add a README file'.
6. Select any license.
7. Click 'Create repository'.
8. You will be automatically directed to your new repository webpage.
9. Now we will get a DOI from the Zenodo application. Note that we are going to use https://sandbox.zenodo.org/ to do this. This offers all the same capabilities as https://zenodo.org but is a testing site! Create a free account if you have not already.

Part 2: Create an archived repository and affiliated DOI.

1. Navigate to the Zenodo GitHub page. Click on the button 'Connect' to allow Zenodo to access your GitHub repositories.

2. Review the information about access permissions, then click 'Authorize Zenodo'.

3. Sync your GitHub with Zenodo by clicking 'Sync now' in the upper right corner.

4. To the right of the name of the repository you want to archive ('os-test'), toggle the button to On.
5. Click on the name of the repository.
6. Click the big green button that has 'username/os-test'

7. Add a tag 'test'. You may have to create a new tag for 'test' if prompted.
8. Scroll down and click the green 'publish release' button

9. Navigate to the Zenodo GitHub page and see the DOI for 'os-test'
10. Share your DOI below.

Zenodo archives your repository and issues a new DOI each time you create a new GitHub release. Follow the steps at "Managing releases in a repository" to create a new one.

20.8.4 Making it Easy to Cite Your Code

Information about how to cite the software can then be added to your README or other documentation in your repository. Another useful step for making your repository citation information accessible is to add a CITATION file to the repository.

20.8.5 Why use CITATION files?

CITATION files are a means to make citation information easily accessible in open source software repositories. A citation file format (CFF) is a human and machine-readable standard format that has been developed for CITATION files.

20.8.6 Adding Contributor Guidelines

If you are hoping for community input on your software, it is a best practice to include CONTRIBUTING and CODE_OF_CONDUCT files in your repository that outline expectations for member interactions.

We won't go into these in detail here, but you can check out the Xarray package's github repository for a good example.

20.9 Who: Roles and Responsibilities of the Team Members in Implementing the SMP

When writing a SMP, it's important to include a plan for the roles and responsibilities needed to share and (if applicable) maintain your code. Your community will consist of members in different roles – some actively engaged, some with only a passing interest. Sometimes, multiple roles can easily be done by one person (e.g. if you are just archiving a piece of code).

Some roles might include:

Who will add the code to a public repository? - Uploading the code - Assigning a license

Who will take care of code documentation - Writing a README - Adding explanatory comments to the code

Who will help with code reuse?

Adding CITATION, CONTRIBUTING, and CODE_OF_CONDUCT files

Who will maintain the software (if applicable)? - Who will respond to community input (e.g. via GitHub issues)? - Who will be responsible for making decisions about which code to add/update from other contributors? (e.g. via GitHub pull requests)

All of these roles may or may not be needed, depending on the size of your project. Have a transparent process for assigning any roles to community members.

20.9.1 Responsibilities after Sharing

If the software is meant for others to use, then the developer should maintain the software.

- It is polite for the developer to let users know whether or not they intend to maintain the software/code.
- Do this in the documentation where you discuss the development status of the project.
- This will help users know if it will continue to be supported in the future and allow them to make choices about basing ongoing work off your project.
- In the case that a developer/researcher may not have the time or continued funding to keep up with a project but others are interested in keeping it maintained, consider handing ownership of the software to another researcher/developer, involved user or entity invested in its continued use.

- Users of software that is no longer maintained may consider contacting the owner/developer and volunteering either as a maintainer or to take over ownership of the project.
- If you decide to maintain your software, you should respond to requests for features and fixes as you are able.

20.10 Lesson 4: Summary

In this lesson, you learned the key steps in sharing open software:

- Should you share? When sharing software, the policies of your institution and funding agency must be followed. These may limit the openness of the software. Software sharing policies also vary by organization.
- When to share? Follow guidance from your organization, funding agency, or publisher.
- Where to share? It depends on whether you are archiving or sharing for community input. Use domain-specific repositories where appropriate.
- How to enable reuse? Enable reuse through assigning a DOI, and include a license, citation information, and contributor guidelines.
- Who helps share? Plan for the roles and responsibilities when sharing and (if applicable) for maintaining software.

20.11 Lesson 4: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/06

Read the statement and decide whether it's true or false:

I don't need to share my code if I don't plan to continue developing it.

- True
- False

Question

02/06

Read the statement and decide whether it's true or false:

Adding code to a GitHub repository is sufficient for archiving my code.

- True

- False

Question

03/06

Read the statement and decide whether it's true or false:

Organization and government software-sharing policies follow a standard practice.

- True
- False

Question

04/06

Read the statement and decide whether it's true or false:

Publishing your software to a software repository used by common package managers makes it easier for users to install your software.

- True
- False

Question

05/06

Which, if any, of the following are ways you can help others to reuse your code? Select all that apply.

- Assign an appropriate license
- Add a file named "CONTRIBUTING" with contributor guidelines
- Add a "CITATION" file with citation information

Question

06/06

Which of the following are roles that you should plan for when writing a SMP? Select all that apply.

- Who will help maintain the software
- Who will create the repository and add the necessary files
- Who will contribute to the software after it is shared
- Who will add documentation to the software

21 Lesson 5: From Theory to Practice

21.1 Navigation

21.2 Overview

This lesson ties the concepts of open access software development to the operation of a software management plan. The lesson also introduces you to the community aspect of open software. It begins with a discussion on writing software management plans, then continues with information on how to connect with open software communities. This information is contextualized with an introduction to the benefits of a software community and the roles involved in these groups. A list of communities is also presented, and you are asked to explore and engage with some of them. The lesson wraps up with helpful suggestions to contribute to open software and additional resources.

21.3 Learning Objectives

After completing this lesson, you should be able to:

- Recall the definition of a software management plan, potentially as part of an open science and data management plan, and where to find helpful resources.
- List ways to engage with and contribute to open software communities.

21.4 Open Science and Data Management Plans

"A NASA Open Science and Data Management Plan (OSDMP) describes how the scientific information that will be produced from NASA-funded scientific activities will be managed and made openly available. The OSDMP should include sections on data management, software management, and publication sharing."

https://science.nasa.gov/researchers/sara/faqs/

Example sections to include in an OSDMP:

- Data Management Plan (DMP)
- Software Management Plan (SMP)
- Publication sharing
- Other open science activities
- Roles and responsibilities

21.4.0.1 Recall the steps for an SMP from the previous lessons

- **What:** Description of management, preservation, and release of software.
- **When:** The schedule for software archiving and sharing.
- **Where:** Location where software will be shared and archived over the long-term.
- **How:** Enable reuse of software through assigning a DOI, license, contribution guidelines, etc.
- **Who:** Roles and responsibilities of the team members.

21.5 How Do We Plan for Making our Code Open?

21.5.1 Should a Software or Data Management Plan be Written?

If you are planning a project that requires a data management plan, writing that plan is a good first step. There is a threshold above which you should write a software/data management plan. "Software" here means scientifically or technically relevant computer programs as both source code and executable software.

SMP is required	You need a SMP to:
SMP is not required	You probably don't need a SMP if you are working on:

Perhaps your project does not fit into these categories. For example, if your aim is for your results to be reproduced by others then writing a SMP is your discretion.

The following material assumes that you have met the threshold and are writing a data/software management plan.

21.5.2 Pen to Paper: Getting Started Writing a Plan

If you are applying for funding, it is almost guaranteed that there will be specific data management requirements detailed in the funding opportunity. For example, the funder may require a certain license or use of a specific repository. Make sure to cross reference your plan with these requirements.

Examples of Software Management Plans

- software.ac.uk/resources/guides/software-management-plans
- software.ac.uk/software-management-plans
- esciencecenter.nl/national-guidelines-for-software-management-plans/
- https://zenodo.org/record/7589725

Policies

What are the policies for a SMP? (what does the funding agency say to do?) - Data formats - Plan for data/code archival/preservation - Roles and responsibilities

21.5.3 Funding Agencies

Scientific funding agencies generally solicit peer reviews to support funding decisions. These reviews explicitly or implicitly evaluate related open software. Community participation is necessary to arrive at consensus regarding community standards for funding.

For example: NASA policy explicitly states that "funded software should follow best practices in the relevant open source and research communities."

21.5.4 Established Open Software Policies of Professional Societies

Professional societies such as AAAS, AGU, AAS, etc., influence funding agency policies and directly influence the policies surrounding software used to generate publications. It is important to engage with the community via consensus papers and professional societies to guide policy decisions regarding open source software in science.

Science/AAAS explicitly states that "In general, all computer code central to the findings being reported should be available to readers to ensure reproducibility."

21.5.5 Institutions

The individual institutions where we work impose highly variable restrictions on open source software due to security, privacy, intellectual property, commercial, or other concerns which do not necessarily align with the ethos of open science. It is important to engage with the institutional community to facilitate the movement toward policies that facilitate open source software as a foundation of open science.

21.5.6 Activity 5.1: Writing an SMP

In this activity, review the SMP below and think about these questions:

- What kinds of software does the SMP describe?
- When will it be shared?
- Where will it be shared?
- How will it be shared so it is a citable artifact?
- Who will be responsible for different aspects of the software?
- What are some of the limitations for some of the software?
- How does not having an agreed upon plan when you start code development have impacts years down the line?
- Are results reproducible without the original IDL code?
- Are there things in the example plan that you would add or be more specific about?

21.5.6.1 Example Software Management Plan

1. Expected Software Types

We will use established simulation models to conduct initial simulations for this work. These simulation models are written in Fortran and developed over the last decade. While not publicly available, they are available for the project to use (private communication). The simulation models will lead to the generation of output files as described in the Data Management Plan (DMP). We will develop analysis software in Python to analyze the model output files, which will enable the development of derived data products, maps, and figures. Development of the Python analysis software will be shared on a GitHub repository.

2. Development of Analysis Software

All new development of Python code will be conducted openly on GitHub by members of this project. We will post and follow the established Code of Conduct for software development for our research project, which includes guidelines for contributions by additional members of the scientific community.

3. Repositories and Timeline for Sharing Software

This work will support the development of two peer-reviewed journal articles. All source code developed in Python to support each article will be archived on Zenodo no later than the article's publication date. The software will be made available under a permissive Apache License 2.0. Zenodo will assign a DOI to the archived software when it is archived.

4. Software Sharing Exemptions

This work does not support further development of the existing Fortran simulation models, which are maintained independently. We do not have permission to publicly share the Fortran source code for the simulation models.

5. Roles and Responsibilities

Initial simulation modeling and the development of Python analysis software will be completed by PhD students and postdocs. The PI of this project holds overall responsibility for the execution of this plan.

21.6 Engage and Build Communities

Open software communities are social learning spaces where individuals come together to learn a new skill, exchange knowledge and experiences, and then apply what they've learned from the community in their day-to-day work.

21.6.0.1 Communities offer:

- A low entry point for learning and improving your use of software in research.
- An opportunity to share individual experiences, identify common hurdles, and iteratively enhance knowledge and resolve problems.
- A way to build the culture around open source software in science and a great way to keep updated on the latest tools and practices.
- A non-hierarchical community of practice where all members of the community should be treated equally.

21.6.1 Connect with Communities

Here are some communities that can help you get started:

- PyData
- SPEC
- rOpenSci
- pyOpenSci
- PyHC

- Research Software Engineering
- NumFOCUS
- R-Ladies
- PyLadies
- WoCCode
- Pangeo
- ObsPy

Subscribe to and/or participate in forums (e.g., GitHub discussions, Stack Overflow, or discipline/software specific), in-person workshops, conferences, hackathons, etc., related to your discipline or software you contribute to or use. Connect on social media. And last but not least, talk with your colleagues!

Explore: The Turing Way

Hit the button to find out more information on building a community.

CLICK TO LEARN

21.6.2 Activity 5.2: Browse Through Some of the Communities of Practice

- Find and browse through the websites associated with two communities of practice listed on the previous section "Connecting with Communities".
- Identify at least two points of entry for engagement, e.g., an upcoming event (virtual or in person), how you could contribute, forums, etc.

21.6.2.1 Key takeaways: Browse through some of the communities of practice

- There are many opportunities to engage with communities working on open software.
- Engaging with open software communities can enrich and improve your software.

21.7 Contribute to Open-Source Software

Contributing to open software provides many advantages and opens doors to a number of rewarding opportunities. There are few other industries that can boast the massive number of global contributions like the open-source community can. Contributing to open source software is a great way to improve your coding skills and to document your work while growing your community.

There are several types of contributing to open software. Not all of them require writing actual code:

Add New Features	The most obvious case for contributing to open software is enhancing its usability by adding new features.
Fix Bugs	Alternatively, you can reply to an already opened issue by fixing it.
Report Issues and Make Suggestions About Improving Code	Reporting an issue is a valuable contribution even if you don't know how to fix it. For example, you might be using a different browser in which the software has not been tested yet, have discovered a particularly uninformative error message, be colorblind or be otherwise able to feed a valuable user experience back to the developers that can help to improve the overall usability of the software.
Improving and Contributing to Documentation	Contributing to documentation constitutes a great starting point to contributing to open source software and is often overlooked in its importance. Writing documentations allows you to familiarize yourself with the use of the software, while helping to teach others.
Create Tutorials, Use Cases, or Visuals	Another way to contribute is to make your experience and use of the software publicly available. For example, you could create a tutorial based on your use of the software, summarize a use case or provide a summary of your use in a graphic. This part of contribution is particularly appealing as it does not create much extra work to just publish what you have used the software for.
Improve Layout, Automatization, Structure of Code	Apart from creating new code, a good way to contribute to open source software can also be to improve, restructure or automatize existing code. This is called refactoring and helps to make the software project more effective and stable.

Organize and Attend a Community Meet-Up	Another way to contribute to open source software is via community building. Many software products and toolboxes have a lively community of users that meet on a regular basis in person and online to discuss and improve the software and its use. Participating or even organizing such a meetup can be a good way to improve your knowledge of the software, get to know its community, and contribute to open source projects.
Code Review	Requests to integrate new contributions into the main code base usually require a review of the contribution by at least one other user. Similar to peer review, code review entails writing a short summary about the quality of the code and making suggestions about improvements.

21.8 Additional Resources

21.8.1 References and Guides

In addition to the resources listed elsewhere in this training, the below community resources are excellent sources of information about Open Software.

- OpenSciency
- NASA SMD's Open-Source Science Guidance
- Practical Guide to Software Management
- FAIR Principles for Research Software (FAIR4RS Principles)
- Open Source Software Policy Options for NASA Earth and Space Sciences
- Turing Way handbook to reproducible, ethical and collaborative data science
- Ten simple rules for documenting scientific software
- Journal of Open Source Software

21.8.2 Additional Training

In addition to the resources listed elsewhere in this training, the below resources represent additional training on Open Source Software.

- Software Carpentries
- How to contribute to Open Source projects – A beginners guide

21.8.3 A Journal with Thousands of Open-Source Research Software Success Stories

The Journal of Open Source Software has presented a venue for enhancing the quality and minimizing the effort of publishing open source research software:

- Peer-reviewed, open source "journal" covering open source research software published via GitHub.
- The emphasis is on the software.
- Published thousands of open source research software projects, several of which are highly cited. JOSS is one of several journals. Click here for a list of many more journals that publish software.

21.9 Lesson 5: Summary

In this lesson, you learned:

- When a SMP should be written and that your funding organization or institution may have rules around how you develop and share your code.
- That joining software communities can be a great way to exchange knowledge and learn new skills around open code.
- That there are many ways to contribute to open code, and that not all of them require writing code."

21.10 Lesson 5: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/02

Read the statement and decide whether it's true or false:

Community engagement with open software is non-hierarchical; all members of the community should be treated the same.

- True
- False

Question

02/02

Select the beneficial way(s) to contribute to open sources software.

- Add new features
- Fix bugs
- Document your work
- Refactoring
- All of the above

21.11 Open Code Summary

Congratulations! Now you should be able to:

- Explain what open source software means, including the software development cycle, the benefits of open software, and some common limitations and how they are addressed.
- Discover open source software and assess it for reuse by evaluating provided documentation, including README files and licensing details; cite the software when appropriate.
- Create an open source software management plan that includes the strategy for selecting open software dependencies and open repositories such as GIT, and how open elements including metadata, README files and version control, will be included to make the software reusable and findable.
- Evaluate whether your open source software can be shared and the best options for sharing to increase visibility.
- List the responsibilities a software developer has once the open source software is shared including managing legal requirements and ensuring the software is maintained.

Part V

OS101 Module 5: Open Results

Welcome to Open Science 101: Open Results

About This Module

Welcome to Open Results! This module focuses on giving you the tools you need to kick-start a scientific collaboration by creating contributor guidelines that ensure ethical contributorship. It starts out with a use case of open science in action, then a review of how to discover and assess open results. Next, the focus is on how to publish results which includes a task checklist. The module wraps up with specific guidance for writing the sharing results section of the Open Science and Data Management Plans (OSDMP). We will also reflect on how our society and technology are constantly evolving in the way we do science.

Module Learning Objectives

After completing this module you should be able to:

- Describe what constitutes an open result.
- Explain what the reproducibility crisis is and how open science can help combat it.
- Use a process to discover, assess, and cite open results for reuse.
- List the responsibilities of the following participants that are creating open results: open results user, project leader, collaborator, contributor and author.
- List the tasks for creating reproducible results and the items to include in a manuscript to ensure reproducible results.
- Define a strategy for sharing your results including selecting publishers, interpreting journal policies and licenses, and determining when to share your data or software with your manuscript.

Part VI

Key Terms

These key terms are important topics for this module. Select the term to see the description.

Research Object (RO) - A method for the identification, aggregation, and exchange of scholarly information on the Web. This can include publications in different forms, software, data, and media.

Predatory Publishing - For-profit publishers that charge a publishing fee but provide few quality checks on the quality of the publication that would be expected from scholarly publications such as peer review or type-setting.

Preprint - A version of a paper prior to the publication in a journal. This can be the author's version of the accepted manuscript after peer review or a version prior to submission to a journal.

Preregistering - A practice by researchers who determine their analysis plan and data collection procedure before a study begins.

Persistent Identifiers - Long lasting reference to a document, file, web page, or other digital object. It is usually used in the context of digital objects that are accessible over the internet. Most PIDs have a unique identifier which is linked to the current address of the metadata or content.

Reproducibility Crisis - The 'reproducibility crisis' in science is a growing concern over several reproducibility studies where previous positive results were not reproduced.

DOI - A digital object identifier is a persistent identifier's handle used to uniquely identify various objects, standardized by the International Organization for Standardization (ISO).

Code of Conduct - A collection of rules and policies that outline the standards, principles, expectations, and morals for a particular group or organization. It is considered binding on any person who is a member of that group or organization. It can help employees align their behavior with the company's values, support decision-making, and foster retention and loyalty.

Navigation

Lesson 1: Introduction to Open Results

Lesson 2: Using Open Results

- How to Discover Open Results
- How to Assess Open Results
- How to Use Open Results
- How to Cite Open Results
- Lesson 2: Summary
- Lesson 2: Knowledge Check

Lesson 3: Making Open Results

- How to Make Open Results
- Role of Contributors in Open Science
- How to Give Open Recognition
- Combining Open Results for Scientific Reporting and Publications
- Lesson 3: Summary
- Lesson 3: Knowledge Check

Lesson 4: Sharing Open Results

- When to Share
- How to Share
- Other Considerations When Sharing
- Lesson 4: Summary
- Lesson 4: Knowledge Check

Lesson 5: From Theory to Practice

- Writing an OSDMP: What to Include in the OSDMP for Sharing Results Openly
- Example Steps Toward More Open Results
- How Emerging Technology Like AI is Changing How We Do Science
- Lesson 5: Summary
- Lesson 5: Knowledge Check
- Open Results Summary
- Open Science 101 Summary

22 Lesson 1: Introduction to Open Results

22.1 Navigation

22.2 Overview

This lesson aims to broaden your perspective regarding what shareable research outputs are produced throughout the research lifecycle. We will first consider what constitutes an open result. To do so, we will read an example of a forward-thinking research project that utilizes open result best practices. The perspectives gained from this example will ultimately get us thinking about how we can work towards creating reproducible research.

22.3 Learning Objectives

After completing this lesson, you should be able to:

- Describe what constitutes open results and list the research objects that can be created throughout a research cycle.
- Describe how sharing open results can advance science and your career.
- Explain what the reproducibility crisis is and how open science can help combat it.

22.4 What Research Objects are Created Throughout the Research Cycle?

22.4.1 The Traditional Depiction of a "Scientific Result" Has Changed Over Time

When we think of results, most people think of just the final publication.

1665

This publication dates back to 1665 when the first scientific journal Philosophical Transactions was established to publish letters about scientific observations and experimentations.

1940s

Later in the 1940s, publishing became commercialized and took over as the mechanism for releasing journals, conference proceedings, and books. This new business model normalized publication paywalls.

21st century

Only by the 21st century did the scientific community expand the meaning of open results. The evolution of this definition was driven by technological advances, such as the internet, and advances in modes to share information. The open access movement was established by the Budapest Open Access Initiative in 2002 and the Berlin Declaration on Open Access in 2003, both of which formalized the idea that, with regards to new knowledge, there should be "free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles" (Budapest Open Access Initiative).

22.4.2 But Results Have Always Been Far More Than Just the Publication

You might be familiar with the research life cycle, but may not have considered what results could be shared openly throughout its process. This lesson adopts a definition of the research life cycle based on The Turing Way and breaks it down into nine phases based, pictured in the figure below.

Although the phases are presented in a linear fashion, we acknowledge that the research life-cycle is rarely ever linear! Products are created throughout the scientific process that are needed to enable others to reproduce the findings. The products of research include data, code, analysis pipelines, papers, and more!

Following Garcia-Silva et al. 2019, we define a Research object (RO) as a method for the identification, aggregation and exchange of scholarly information on the Web. Research objects can be composed of both research data and digital research objects that are defined as follows by the Organization for Economic Co-Operation and Development (OECD Legal Instruments).

The term 'Open Results' comprehensively includes all these research products and more.

Open results can include both data and code. Since data and code were covered in previous modules, in this lesson, we focus on sharing science outcomes as open results. Examples of open results can include:

- Open access peer-reviewed articles
- Technical reports
- Computational notebooks
- Code of conduct, contributor guidelines, publication policies
- Blog posts
- Short form videos and podcasts
- Social media posts
- Conference abstracts and presentations
- Forum discussions

Open access peer-reviewed articles are archived for long-term preservation and represent a more formal discussion of scientific ideas, interpretations, and conclusions. These discussions inform the method that researchers share results. In the following lesson section, we will discuss different types of sharing and methods to build and adapt them for use in your research.

Scientists can share their incremental progress throughout the research process and invite community feedback. Sharing more parts of the research process creates more interactions between researchers and can improve the end result (which may be a peer-reviewed article).

Throughout this module we will show you how to use, make, and share open results.

22.4.3 The Practice of 'Open'

Specifically, the "Use, Make, Share" format has been naturally embedded throughout the curriculum and should be a familiar format by now. Lesson 2 will cover "Using". Lesson 3 will cover "Making". Lesson 4 will cover "Sharing". Throughout this module, we will pay particular attention to manuscripts and other research products as examples because the previous modules covered "Use, Make, Share" in the context of components with data and software.

22.5 Examples of Open Results

Let's broaden our perspectives on the types of **research objects** that are produced throughout the research process. Let's take a look at some examples from different projects.

22.5.1 Reaching New Audiences

Qiusheng Wu is an associate professor at University of Tennessee. He has published 500+ video tutorials on YouTube, which have gained 25K+ subscribers, and 1.1M+ views (as of 8/2023).

Professor Qiusheng Wu created a YouTube channel in April 2020 for the purpose of sharing video tutorials on the geemap Python package that he was developing. Since then, Wu has published over 500 video tutorials on open-source geospatial topics. The channel has gained over 25K subscribers, with more than 1 million views and 60K watch hours in total. On average, it receives 70 watch hours per day.

The YouTube channel has allowed Wu to reach a much larger audience beyond the confines of a traditional classroom. It has made cutting-edge geospatial research more accessible to the general public and has led to collaborations with individuals from around the world. This has been particularly beneficial for Wu's tenure promotion as it has resulted in increased funding opportunities, publications, and public engagement through the YouTube channel, social media, and GitHub.

Overall, the YouTube channel serves as an important tool for Wu to disseminate research, inspire others, and contribute to the advancement of science. It has also played a significant role in advancing Wu's professional career.

22.5.2 New Media for Science Products

"A new method reduced the compute time for this image from ~30 minutes to <1 minute". In 2021, Lucas Sterzinger spent one summer of his PhD on an internship. During that summer, he wrote a blog post to explain and demonstrate a game-changing technology called Kerchunk – a software package that makes accessing scientific data in the cloud much faster.

Source: https://medium.com/pangeo/fake-it-until-you-make-it-reading-goes-netcdf4-data-on-aws-s3-as-zarr-for-rapid-data-access-61e33f8fe685

Alongside the blog post, he also created a tutorial as a Jupyter Notebook – both of these resources and associated code are freely accessible to the public, allowing for rapid adoption and iteration by other developers and scientists. He posted the blog on Medium and posted about it to Twitter. The blog got a lot of attention on a newly developed technology as it was being developed! This is starkly different from the slow and complicated world of academic publishing where this result would not have been shared for about a year (writing it up, the review process, publication process). He said, "Working on Kerchunk and sharing it widely using open science principles greatly expanded my professional connections and introduced me

to the field of research software engineering. The connections I made from this led me directly to my current role as a Scientific Software Developer at NASA."

22.5.3 New Products for Increasing Impact

Image credit: OpenStreetMap 2011, Ken Vermette. CC BY-SA 3.0

From '2003: let's map the UK to 2023:>1.5M contributors, 100M+ edits, using the data to map the world with applications ranging from Uber to mapping UN Sustainable Development Goals." OpenStreetMaps is being used for GIS analysis, such as planning or logistics for humanitarian groups, utilities, governments and more. This was only possible because it was set up and shared openly and built by a community devoted to improving it. You never know where your personal project might go or who might be interested in collaborating!

22.5.4 New Visualizations to Share Results

Matplotlib was developed around 2002 by post-doc John Hunter to visualize some neurobiology data he was working on. He wasn't a software developer, he was a neurobiologist! He could have just published the paper in a peer-reviewed journal, and maybe shared his code to create the figures, but instead he started an open project on GitHub and thought, 'well if this is useful to me, maybe it will be useful to others…'.

Source: https://medium.com/dataseries/mastering-matplotlib-part-1-a480109171e3

Matplotlib is now the most widely used plotting library for the Python programming language and a core component of the scientific Python stack, along with NumPy, SciPy and IPython. Matplotlib was used for data visualization during the 2008 landing of the Phoenix spacecraft on Mars and for the creation of the first image of a black hole.

22.5.5 JWST Case Study: Reporting and Publication

And last but not least, we have the example for the JWST Early Release Science team from Module 1 on how they reported their results. This came in various forms from publishing a peer review paper, preprints, blog posts, and social media. Their peer-reviewed publication was published open access in Nature along with a preprint through arXiv.

Open communication platforms furthered the reach and audience of results.

Figure Credit: https://arxiv.org/abs/2208.11692

The public is interested in what you are doing, and reaching them can involve communication through traditional and new platforms. Publishing results on platforms such as Twitter/X, Youtube, TikTok, blogs, websites, and other social media platforms is becoming more common. Awareness through social media drastically increases the reach and audience of your work. There have been studies on how this impacts citation rates. For example, The Journal of Medical Internet Research (JMIR) conducted a three-year study of the relative success of JMIR articles in both Twitter and academic worlds. They found that highly tweeted articles were 11 times more likely to be highly cited than less tweeted articles.

Open communication platforms noticeably furthered the reach and audience of results.

Twitter #1: https://twitter.com/cornerof_thesky/status/1595086671275589632?s=20

Twitter #2: https://twitter.com/V_Parmentier/status/1595127493199302656?s=20

TikTok: https://www.tiktok.com/@astrojaket/video/7168878696906886405

YouTube: https://www.youtube.com/watch?v=cI-kM_wPbbQ

22.6 What is the Reproducibility Crisis?

A 2016 Nature survey on reproducibility found that of 1,576 researchers, "More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments."The 'reproducibility crisis' in science is a growing concern over several reproducibility studies where previous positive results were not reproduced.

We must consider the full research workflow if we are to solve the reproducibility crisis. The fact that 70% of researchers could not reproduce other scientists' results is shocking, especially considering that the reproducibility of science is the cornerstone of the scientific method.

By now, it should be obvious that there are many personal incentives to implement open science principles throughout all stages of the research process. By making results open throughout, you increase your ability to reproduce your own results.

Although reproducibility of one's own results might sound like a trivial achievement, a 2016 Nature study found that 50% of researchers are unable to reproduce their own experiments. This highlights the critical nature of the reproducibility crisis. This also has implications for research beyond the ability to improve your research.

22.6.1 What is the Cause of This Reproducibility Crisis?

The three main causes of the reproducibility crisis are:

1. Intermediate methods of research are often described informally or not at all.
2. Intermediate data are often omitted entirely.
3. We often only think about results at the time of publication.

We need to think of the entire research process as a result. As an example, scientific articles describe computational methods informally which demands significant effort from others to understand and to reuse.

Articles often lack sufficient information needed for other researchers to reproduce results, even when data sets are published, according to two studies in Nature Genetics and Nature Methods. Raw and/or intermediate data products and relevant software are often not provided alongside the final manuscript, limiting the reader's ability to attempt replication.

Without access to the source codes for the papers, reproducibility has been shown to be elusive, according to two other studies in Briefings in Bioinformatics and Nature Physics.

22.6.2 Combating the Reproducibility Crisis

If your research workflow uses principles of open results, as showcased in the example, this will help you to combat the reproducibility crisis.

We can create reproducible workflows and combat this crisis by considering open results at each stage of the research lifecycle. An Open Science and Data Management Plan (OSDMP) helps researchers think and plan for all aspects of sharing by determining how they will make software and data available. This plan can be shared publicly early on through a practice called pre-registering, where researchers determine their analysis plan and data collection procedure before a study begins (discussed previously in Lesson 2 of Module 2).

22.6.3 Activity 1.1: What Could You Do?

Let's rethink your research workflow. Identify the research objects that could be (or could have been) shared as open results of a project you are/were involved in. What are high priority items for combatting the reproducibility crisis in each area of the research workflow?

- Ideation
- Planning
- Project Design
- Engagement & Training
- Data Collection
- Data Wrangling
- Data Exploration
- Preservation
- Reporting & Publication

There are many personal advantages of implementing open science principles across all stages of a research process

22.6.3.1 Key Takeaways: What Could You Do?

The OpenSciency team created a large table that describes all the different kinds of shareable research objects that are possible to create throughout the research lifecycle.

A full table is available here

CLICK TO LEARN

Thinking about sharing everything all at once can be overwhelming when you are getting started. To move forward, just focus on how you might pick the most important item. Here we have pared down the list to only a couple items per category. Furthermore, you could think about shortening the list even further when you are getting started. For example, maybe it is the case that, for your work, sharing the code used to wrangle the data is the most critical element to reproducibility. Therefore, code-sharing would be a good place to start your open science journey. The small steps we make are what move us towards sustainable open science.

- **Ideation:** Proposals can be shared on Zenodo and open grant platforms such as ogrants.org.
- **Planning:** Projects can be pre-registered before they begin.
- **Project Design:** Contributor guidelines or a code of conduct can be posted on Zenodo, GitHub, or team Web Pages.
- **Engagement & Training:** Workflow computational notebooks can be shared with the team via GitHub and released on Zenodo.

- **Data Collection:** Raw data can be shared through data repositories.
- **Data Wrangling:** Code can be shared through software repositories.
- **Data Exploration:** Computational notebooks can be shared via GitHub and released on Zenodo.
- **Preservation:** Data management plans for archiving can be posted on Zenodo.
- **Reporting & Publication:**
  - Open access peer-reviewed articles
  - Computational notebooks
  - Code of conduct, contributor guidelines, publication policies
  - Blog posts
  - Short form videos and podcasts
  - Social media posts
  - Conference abstracts, posters, and presentations (when made openly available)
  - Forum discussions

22.7 Lesson 1: Summary

In this lesson, you learned that:

- The contemporary scientific workflow involves being open about processes and products. Research products (results) include far more than just the final manuscript, which is a drastic change from the historical notion of a scientific result.
- At every stage of the research lifecycle, there are research objects produced that we can consider results.
- We can combat the reproducibility crisis by sharing these research objects at each stage of our research workflow.
- There are amazing examples of research groups sharing different types of open results!

Let's start thinking about what we can do immediately to work towards an open research workflow.

22.8 Lesson 1: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/02

Which of the following may fit the definition of a "research object"?

- Raw data

- Blog
- Proposal
- Code of Conduct
- All of the above

Question

02/02

What are some of the key causes of the reproducibility crisis?

- Intermediate methods of research are often described informally or not at all.
- Intermediate data are often omitted entirely.
- We often only think about results at the time of publication.
- All of the above

23 Lesson 2: Using Open Results

23.1 Navigation

23.2 Overview

By the end of this lesson, you will be familiar with resources for open results utilization, how and when to cite the sources of the open results that you use, how to provide feedback to open results providers, and how to determine when it is appropriate to invite authors of the open results materials to be formal collaborators versus simply citing those resources in your work.

Published articles, blog posts, and forums can lead to new ideas for your own research. A technique learned from social media can be applied to a use-case that you are trying to solve. There are many different ways to discover results.

23.3 Learning Objectives

After completing this lesson, you should be able to:

- Identify a variety of open results sources including both published science research and non-traditional sources.
- Evaluate the reliability and quality of open results sources based on key characteristics.
- List the responsibilities of an open results user, including providing feedback to open results developers.
- List the ways to cite open results into your own research process.

23.4 How to Discover Open Results

How do I learn about the state of research for a particular field? How do you engage in the current conversation? Researchers often begin with a search of peer-reviewed articles. This review tells you how much research has been done in a field and what conclusions have recently been reached. In most fields, going through the peer-review process can take up to a year. The ability to find pre-prints can help reduce this delay because they offer the latest findings before a publication date. However, researchers who choose to share their results before publication typically do so in the ways listed as best practices above. As you start research on a topic, how do you find all these different types of results and engage in the most relevant research?

23.4.1 Example: Exoplanets

The various stages of research, from conceptualization to dissemination of results, produce products that can be put into the public domain as "Open Results". Where these results are archived, and to what degree, depends on the discipline author. However, some general guidelines on where to start a search on open results include:

1. Scholarly Search Portals
2. Web Searches

Scholarly Search Portals

Search engines like Google and Bing have radically changed how we look up information. For research results, specialized academic search engines and portals curate scientific results from researchers based on topic and field. These engines are useful for finding peer-reviewed articles.

GENERIC

DISCIPLINE-SPECIFIC

Google Scholar

ADS

Scopus

Web of Knowledge

Open knowledge map - facilitates exploration of interconnected topics

JSTOR - a wide range of scholarly content

ResearchGate

ScienceCast

GENERIC

DISCIPLINE-SPECIFIC

EuropePMC Life sciences

Pubmed biomedical literature

arXiv - for scholarly pre-prints in STEM, economics and computer science fields

Biorxiv Preprint - server for biology

EarthArXiv and Earth and Space Science Open Archive

ASAPbio - catalogs of preprint servers

and others…

Publications that provide some levels of open access are tracked in the Directory of Open Access Journals (DOAJ).

Web Searches

Open results include much more than open-access peer-reviewed publications. How do you find these alternative types of research objects?

Open communities and forums offer the best way to find research objects other than complete publications. How do you even find out whether these exist and where they are?

Once you have found a few peer-reviewed articles that are highly relevant, to find additional research objects, you can follow the authors on social media for links to their posts, blogs, and activities. There are open communities in almost every area of research - find yours! Here are different platforms to locate these conversations and resources:

- GitHub
- LinkedIn
- YouTube
- Google/Bing
- Conference websites
- X, formally known as Twitter
- Facebook
- Medium
- Substack
- Stack Overflow
- Reddit
- Mastodon

Various research objects, including datasets and software, are frequently attached to scholarly publications in the form of supplemental material. At other times, the source is referenced in the paper, which could be a GitHub repository, personal/institutional website, or other storage site. This can be another starting point, by engaging in discussions on the GitHub repository.

Kerchunk Example: In lesson 1, a blog post about a software library 'kerchunk' was presented. Let's look at a post on the Pangeo Discourse Forum of Kerchunk with a large number of views. The open science Pangeo project worked completely in the open. The project website (run off of GitHub) has links to blog posts, a discussion forum, and a calendar to all their meetings which anyone was welcome to join. This has resulted in an engaged and dynamic community. An example of this comes from the post linked to above, where one person asks for help, others reply, and the conversation is documented in the open. The post's 636 views indicate that this question, or one similar, has occurred to others. Imagine if this had been done over private email? By working in the open, they are improving science and helping everyone become faster and more accurate.

23.5 How to Assess Open Results

"Garbage in, garbage out" – your own research products are only as good as the data used in your investigation.

If you use poor quality data or materials from unreliable and unvetted sources as critical components of your research, you run the risk of producing flawed, or low-quality science that may harm your reputation as a scientist. Therefore, it is critical to assess the quality and reliability of open-results sources before you include them in your own work.

What are best practices for assessing the quality of alternative sources of data to research articles such as blog posts, youtube videos, and other research objects?

23.5.1 Attributes of Reputable Material

Let's take a look at the questions you might consider asking yourself when determining the reliability of any type of open results source.

Here, we list questions under two categories: the open results material themselves, and the server they are downloaded from. The more questions here that can be answered in the affirmative, the lower the risk in utilizing the open results materials for your own research.

THE MATERIAL ITSELF

THE ASSOCIATED WEBSITE / SERVER

SOURCE RELIABILITY INDICATORS

Is the material associated with a peer-reviewed publication?

Are the primary data associated with the results also open-source?

Is code used to generate the Open Results materials also open-source?

Are all fields and parameters clearly defined?

Is the derivation of measurement uncertainties clearly described?

Were any data or results excluded, and if so, were criteria provided?

Are authoring teams also members of the field?

THE MATERIAL ITSELF

THE ASSOCIATED WEBSITE / SERVER

SOURCE RELIABILITY INDICATORS

Does the host website's URL end in .edu, .gov or (if managed by a non-profit organization) in .org?

Does the host website provide contact information of the author and/or organization?

Is the host website updated on a frequent basis?

Is the host website free of advertisements and/or sponsored content, the presence of which could indicate bias?

THE MATERIAL ITSELF

THE ASSOCIATED WEBSITE / SERVER

SOURCE RELIABILITY INDICATORS

Is the result reproducible? Can you interact with the data and results? Have others reported being able to reproduce the results?

Is the author reliable? Have you seen them publish or share results in other forums?

Is the result from only a single author/voice or includes contributions from a broader community?

Does the post have a significant amount of likes/views and public comments? The value of a blog post with no comments or responses can be difficult to assess. Conversely, a thorough github discussion forum with multiple views shared indicated a robust post.

Is the result part of an active conversation? (Is the information still relevant and current?)

Adapted from https://www.scribbr.com/working-with-sources/credible-sources/

Note that failure to meet one or many of the criteria does not automatically mean that the open results are of poor quality, but rather that more caution should be exercised if incorporated

into your own research. It also means that you will have to invest more personal vetting of the material to ensure its quality is sufficient for your purposes.

Reliable Example: Qiusheng Wu YouTube videos (as mentioned in the previous lesson). Professor Wu is an expert in his field. He presents results along with notebooks that demonstrate reproducibility. Comments on his YouTube tutorial videos represent meaningful interactions between users reproducing results and the author.

23.6 How to Use Open Results

While open results benefit science and have already provided valuable societal benefits, the misuse and incautious sharing of open materials can have far-reaching harmful effects. The end-user of open results bears the responsibility to ensure that the data they reference are used in a responsible manner and that any relevant guidelines for the use of the data are followed.

23.6.1 How to Contribute and Provide Constructive Feedback

Contributing to and providing constructive feedback are vital components for a healthy open access ecosystem, ensuring long-term sustainability of the open resources by providing continual improvements and capability expansions.

In our current system, there are results creators and consumers. This scenario presents a one way street with no feedback loop, no sharing of data back to publishers, and no sharing between intermediaries.

The practice of producing open results aims to foster a system where feedback loops exist between users and makers. Users share their cleaned, integrated, or improved work to the maker. This feedback creates a symbiotic and sustainable process where everyone benefits.

23.6.2 Your Responsibilities as an Open Results User

- Users should familiarize themselves with contributor guidelines posted to open result repositories and follow the associated policies. What if there aren't contributor guidelines? Contact the creators!
- Always provide feedback in a respectful and supportive manner.
- If you discover an error in Open Results materials, the ethical action to take is to contact the author (or repository, depending on the nature of the issue) and give them the opportunity to correct the problem, rather than ignoring the issue or (worse!) taking advantage of a fixable issue to elevate your own research.

23.6.3 Different Ways to Provide Feedback

23.6.3.1 Use Github Issues

Pro: The feedback is open and other community members can see ongoing issues that are being addressed.

Pro: Contribution is archived and logged on GitHub.

Working with GitHub Issues

See this blog for general issue etiquette

OPEN

23.6.3.2 Email authors

Con: the feedback is closed. The information is generally not propagated back to the community unless the creator creates a new version.

Con: No way of tracking credit.

23.6.4 Getting Credit for Providing Feedback

If your feedback results in a substantial intellectual contribution to the work, it is reasonable for you to expect an opportunity for co-authorship in a future version of the open result. The associated contribution guidelines should address this possibility and manage expectations prior to your providing feedback.

Sadly, many times contributor guidelines do not exist and it is not clear what is "substantial".

23.6.5 Open Results User Responsibilities

- **Institutional Security Compliance:** Always download code from an authoritative source and be familiar with / follow your institution's IT security policies.
- **Licensing Policies:** Understand and abide by the license(s) associated with the open results materials being used.
- **Attribution and Contribution:** Provide appropriate attribution for the open results used and contribute to the open results community.

Additionally, give credit to repositories that provide open source materials in the acknowledgement section of your paper. If the repository provides an acknowledgments template in their "About" link, follow that suggestion. Otherwise, a generic "This research has made use of <insert repository name>." will be sufficient.

23.6.6 Avoid Plagiarism When Using Open Results

Standard guidelines that you've been using in your research all along for providing appropriate attribution and citations of closed access publications also apply to open access published works.

Examples of plagiarism include:

- Word-for-word copying without permission and source acknowledgement.
- Copying components (tables, processes, equipment) without source attribution.
- Paraphrasing an idea without proper source referencing.
- Recycling one's own past work and presenting as a new paper.

23.6.6.1 FACTSHEET: Plagiarism

Here is a useful guide regarding the different forms of plagiarism

CLICK TO LEARN

23.7 How to Cite Open Results

Giving proper attribution to open results is an important and ethical responsibility for using open a source materials. The process for citation is specific to the nature of the material.

23.7.1 Citation Guidelines for Published Versus Unpublished Results

If a paper has been formally published in a journal, then your citation should point to the published version rather than to a preprint server.

Take the time to locate the originating journal to provide an accurate citation.

Preprint Server (Cite only if journal publication not available)

Source Publication (Always cite)

If a paper that you wish to cite is not yet accepted for publication, you should follow the guidelines of the journal to which you are submitting your paper. A preprint reference citation typically includes author name(s), date of the most recent version posted, paper title, name of the preprint server, object type ("preprint"), and the DOI.

At the time of the Lesson preparation, the following paper did not yet appear as a journal publication.

Jin, H., et al. 2023, "Optical color of Type Ib and Ic supernovae and implications for their progenitors," ApJ, preprint, arXiv:2304.10670.

FOR MATERIAL THAT HAS A DOI

FOR MATERIAL THAT DOES NOT HAVE A DOI

FOR OTHER MATERIALS OR INTERACTIONS THAT WERE HELPFUL FOR YOUR RESEARCH

To cite all of the following, follow existing guidelines and community best practices:

Cite publications

Cite data

Cite software

Cite any other object with a DOI. Since many journals will only allow authors to cite material that has a DOI, what do you do with other types of open results?

FOR MATERIAL THAT HAS A DOI

FOR MATERIAL THAT DOES NOT HAVE A DOI

FOR OTHER MATERIALS OR INTERACTIONS THAT WERE HELPFUL FOR YOUR RESEARCH

Examples include blog posts, videos, and notebooks.

You could also contact the author and ask them to obtain a DOI.

Leave a comment in the comments section or on the forum letting the author know about your publication.

FOR MATERIAL THAT HAS A DOI

FOR MATERIAL THAT DOES NOT HAVE A DOI

FOR OTHER MATERIALS OR INTERACTIONS THAT WERE HELPFUL FOR YOUR RESEARCH

Acknowledge communities and forums that helped you advance your research in the Acknowledgements Section. Not only does this give them credit, but it helps others find those communities.

Citing open research results advances science by giving appropriate credit for all parts of the research process. This is essential for the cultural shift to open science; we must give credit for all types of contributions, and expect them in return. Participatory science allows more people, from more places, with different voices and experiences to participate in science.

Contributing and collaborating this way lowers the barriers (like conference fees) to participating in science and broadens who can participate.

23.7.2 Examples of Giving Credit

In the Lesson 1 blog post example, researchers acknowledged people they worked with in an article they wrote that they found helpful, and two different communities, as well as the computational environment they worked on. This is a great example of giving credit: "I would like to thank Rich Signell (USGS) and Martin Durant (Anaconda) for their help in learning this process. If you're interested in seeing more detail on how this works, I recommend Rich's article from 2020 on the topic. I would also like to recognize Pangeo and Pangeo-forge who work hard to make working with big data in geoscience as easy as possible. Work on this project was done on the Pangeo AWS deployment."

In Lesson 1, the JWST case study was presented. The peer-reviewed publication that reported the first discovery of CO2 on another planet has been accessed 18,000+ times. Notice is that the authorship is attributed to the entire team. The Acknowledgements section duly explains the contributions of their collaborators and partners, "The results reported herein benefited during the design phase from collaborations and/or information exchange within NASA's Nexus for Exoplanet System Science (NExSS) research coordination network sponsored by NASA's Science Mission Directorate." Also, "All the data and models presented in this publication can be found at https://doi.org/10.5281/zenodo.6959427". And finally, they cite all the software! "The codes used in this publication to extract, reduce and analyze the data are as follows.."

23.8 Lesson 2: Summary

In this lesson, you learned:

- Open results can be found using both Scholarly Search Portals and Web searches.
- The reliability of a post can generally be evaluated by the trustworthiness of the website from which they originated from, the engagement of community members, and the scientific rigor of its content.

- Users of open results, as inherent stewards of the open a source community, informally carry some responsibility to contribute to the community's sustainability. This participation includes providing feedback to open results providers and developers.
- Giving proper attribution to open results is an important and ethical responsibility for using open source materials. The process for citation is specific to the nature of the material.

23.9 Lesson 2: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/02

Which of the following could be a source of open results? Select all that apply.

- Web searches
- Papers accessed through a paid subscription
- Materials made public after a 1-year exclusive-use/paid-subscription-only period
- Repositories
- Open access papers

Question

02/02

Which of the following characteristics suggest that a particular paper / data set is more likely to be a credible Open Result? Select all that apply.

- The website lists its source of funding
- The results are described in an associated peer-reviewer publication
- Detailed documentation accompanies the data, including defined fields and parameters
- Organization contact information is listed
- The website was last updated in 2015
- The webpage advertises open job positions with the funder
- The author is an expert in the field
- The webpage's URL ends in '.com'
- The accompanying documentation states that "on inspection, data that were obvious outliers were excluded from the following analysis"

24 Lesson 3: Making Open Results

24.1 Navigation

24.2 Overview

In Lesson 2 you learned how to use other's results. In this lesson, we focus on making open results. We will start by discussing what it means to make reproducible results. Having earlier in the course discussed the computational reproducibility practices in open software, in this lesson, we specifically emphasize the importance of collaborations in making those results open and reproducible. This begins with acknowledging that the scientific results are not made by single individuals. We will then teach how to ensure equitable, fair, and successful collaborations when making your open results that acknowledge all contributions. Once you've planned the rules of engagement, we will provide you with ways to ensure that your reporting and publication abide by open results principles and combat the reproducibility crisis.

24.3 Learning Objectives

After completing this lesson, you should be able to:

- Identify approaches to make different types of open results.
- Recognize the importance of collaboration in making results.
- Develop contribution guidelines to enable recognition of contributors who make results.
- Combine different open results to create scientific reports and reproducible outputs.

24.4 How to Make Open Results

24.4.1 Capturing the Research Process Accurately in the Making of Results

> I am aware of the reproducibility crisis and how open science can help combat it. What practical ways can I apply to my research outputs to make open results? How can I ensure that the results I share can be reproduced by others? How can I publish scientific publications that do not add to, but combat the reproducibility crisis?

In the Ethos of Open Science, you learned about the ethics and principles underlying responsible open science practices. In Open Software, you explored and identified the right tools and methods that ensure the usability and reproducibility of your analysis. In Open Data, you developed a data management plan that can ensure the Findability, Accessibility, Interoperability and Reusability (FAIR) of your data throughout the research process, and not just at the end when the final report from the project is released. These open science approaches directly address the root causes of the reproducibility crisis, which are a lack of openness throughout the scientific process, lack of documentation, poor description of intermediate methods or missing data that were used at intermediate stages of the research process. In this lesson, you will learn to put all of these together to ensure that you are prepared to make your open results easy to reproduce by others.

In Lesson 1, we identified different research components that can be considered open results at various stages of research. In this lesson, we want to specifically explain what processes are involved in making them.

24.4.2 Case Study: Open Results from Distributed Multi-Team Event Horizon Telescope Collaboration (EHTC)

Example: Capturing results on activities ranging from collaboration to observations, image generation to interpretation.

In 2017, the Event Horizon Telescope targeted supermassive black holes with the largest apparent event horizons: M87, and Sgr A* in the Galactic Center on four separate days. This distributed collaboration led to the multi-petabyte yield of data that allowed astronomers to unveil the first image of a black hole providing the strongest visual evidence of their existence. The EHTC website provides information about research projects, scientific methods, instruments, press and media resources (such as blog posts, news articles and YouTube videos), as well as events, data, proposals and publications. This project shows large-scale and high-impact work that applies open practices in making their results. Different kinds of outputs shared under this project can be mapped to different stages of the research process and the teams involved in creating them.

24.4.3 Making Results and Crediting Contributors Fairly at Different Stages of Research

The case studies listed above highlight that results associated with a project are more than a publication. By understanding how open results are created in different projects, we can gain deep insights into the processes for making them. With that goal, the rest of this lesson describes the process of making results into three parts: 1) making all types of research outputs; 2) recognizing all contributors; and 3) combining outputs for scientific reporting and publications.

24.4.4 Making All Types of Research Outputs

New ways of working with creative approaches for collaboration and communication in research have opened up opportunities to engage with the broader research communities by sharing scientific outcomes as they develop, rather than at the end through summary articles. A range of research components are created throughout the research lifecycle that can be shared openly. For example, resources created in a scientific project include, but are not limited to the following:

IDEATION AND PLANNING

DATA COLLECTION AND EXPLORATION

COMMUNITY ENGAGEMENT AND REPRODUCIBILITY

PRESERVATION AND PUBLICATION

Ideation and planning – perhaps before the research project is funded or started:

Research proposals

People and organizations involved

Research ethics guidelines

Data management plan

IDEATION AND PLANNING

DATA COLLECTION AND EXPLORATION

COMMUNITY ENGAGEMENT AND REPRODUCIBILITY

PRESERVATION AND PUBLICATION

Data collection and exploration – research artifacts created during the active research process:

Project repository

Project roadmap and milestones

Resource requirements

Project management resources (without sensitive information)

Collaboration processes like Code of Conduct and contributor guidelines

Virtual research environment

Data and metadata information

IDEATION AND PLANNING

DATA COLLECTION AND EXPLORATION

COMMUNITY ENGAGEMENT AND REPRODUCIBILITY

PRESERVATION AND PUBLICATION

Community engagement and reproducibility – most valuable during the project period:

Training and education materials

Computational notebooks

Computational workflow

Code repository (version controlled)

Blog posts

Short form videos and podcasts

Social media posts

Forum discussions (for example when asking for feedback or troubleshooting)

IDEATION AND PLANNING

DATA COLLECTION AND EXPLORATION

COMMUNITY ENGAGEMENT AND REPRODUCIBILITY

PRESERVATION AND PUBLICATION

Preservation and publication – expected to persist long-term:

Publication and authorship guidelines

Open access peer-reviewed articles

Conference abstracts and presentations

End of project report

User manual or documentation

Public outreach and events

Image credit: The Turing Way project illustration by Scriberia. Zenodo.

You have already come across some of these in the previous lessons, and hopefully, you could already identify which of these or additional outputs you are generating in your work. To make them part of your open results, it's important that they are shared openly with appropriate licensing and documentation so that others can read, investigate and when possible, reuse or build upon them.

24.4.5 Making Open and Reproducible Results

Open science ultimately informs our decisions as scientists and guides the selection of approaches that contribute to making our results open at different stages. One of the main purposes of open results is to ensure research reproducibility, often explained through definitions such as the following by Stodden (2015):

"Reproducibility is a researcher's ability to obtain the same results in a published article using the raw data and code used in the original study."

Stodden (2015)

Using this definition, results that can be computationally reproduced by others would be called Reproducible Results. The EHTC case studies present open results as collections of research objects created at different stages of the research process. They also provide documentation and resources that allow reanalysis and reproduction of the original results.

Ideally, anyone, anywhere, must be able to read a publication and understand the results, easily find methods applied, as well as properly follow procedures to achieve the same results as shared in that study. However, as already learned, the issue of reproducibility is prevalent across all scientific fields (refer to this Nature report). A well-intentioned scientist may share all research objects and describe all steps applied in their research, but failing to provide the research environment or other technical setup they used for analyzing their data can prohibit others from reproducing their results. This issue is further compounded by human bias and errors. For example, individuals may not always be able to identify how their interests and experiences inform their decisions that impact their research conclusions. This makes the issue of combating the reproducibility crisis even bigger.

Approaches for making open results should integrate reproducible tools and methods, such as version control, continuous integration, containerisation, code review, code testing and documentation. Furthermore, to extend the reproducibility beyond computational aspects of research, reporting and documentation for different types of outputs and decisions should also be supplied transparently.

24.4.6 How to Make Different Types of Open Results

Sharing different types of results as early as possible not only helps you find solutions faster, but also helps your science be more reproducible because that openness helps you understand how to communicate your methodologies and your findings more clearly to others. Here we provide some easy places to start creating your results openly.

WRITING A FORUM POST

WRITING A GOOD BLOG POST

MAKING A GOOD VIDEO

WRITING A SOCIAL MEDIA POST

Often, when first starting in research, public forums are a great place to begin understanding and collaborating with communities. Most discussion forums have a code of conduct and guidelines on best practices for participation. Some common ones that may be helpful are guidelines from StackOverflow, and Xarray, but most forums have some specific guidance. On forums, you increase trust by interacting with the community, so the more you interact, the more people are likely to respond! Often, best practices include making sure you are posting to the right area, using tags (when available), and including examples that document the question or issue you are having. If you review the post on the Pangeo Discourse Forum with a large number of reviews you can see that they clearly state the problem they are trying to solve, reference other posts on similar topics, link to a computational notebook that has an example of their code, and give an example of the code they are trying to do.

WRITING A FORUM POST

WRITING A GOOD BLOG POST

MAKING A GOOD VIDEO

WRITING A SOCIAL MEDIA POST

Blogs are long-form articles that aren't peer-reviewed. Blogs can be a great way to share your scientific process and findings before they are published, but also after they are published to provide another more accessible presentation of the material. For example, maybe you write a scientific article on your research that is highly technical, but then break it down in more accessible language in a blog post. Many scientists use blog posts to develop and test ideas and approaches because they are more interactive. There are science blogs all over the internet.

Some popular ones are Medium, Science Bites, and Scientific American. One good way to get started is to find a blog post that you liked or found inspirational and use that as a guide for writing your own post.

WRITING A FORUM POST

WRITING A GOOD BLOG POST

MAKING A GOOD VIDEO

WRITING A SOCIAL MEDIA POST

Start small! Record a short video where you show how to do something that you struggled with or a new skill or tool that you learned how to use and post it to YouTube or other popular video platforms. Great videos often explain science concepts, ideas, or experiments to a target audience. Videos can inspire others to work in science, so talk about how you got into science, and show some of your research. There are a lot of online resources to help you out here as well!

WRITING A FORUM POST

WRITING A GOOD BLOG POST

MAKING A GOOD VIDEO

WRITING A SOCIAL MEDIA POST

Social media is also a good place to ask questions as you are just starting on a research topic and also as a place to share all types of results. Providing a link to a video, blog post, or computational notebook and/or sharing an image of a scientific result is a great way to start interactions. You can draw attention to your post by using hashtags and tagging other collaborators. There are a lot of online guides for how to write social media posts and it is always good to look at what others in your area are doing. Responding to comments and engaging with others can help you improve your research and learn about new tools or methods.

All these different ways of sharing information will help make your published report or article better. And as you start working more in the open, with others, think about how collaborations will work and how you will give credit. All resources can be centralized through reports and documentation on a repository or website so anyone, including the 'future you' can find them in the future.

More ways to communicate your work can be found in a guide for communication in The Turing Way.

24.4.7 Maintaining Ethical Standards

Open science, as learned in the Ethos of Open Science, should maintain the highest ethical standards. This can be enabled through the involvement of diverse contributors in the development of scientific outcomes. Participatory approaches allow multiple perspectives and expertise to be integrated into research from the start and ensure that peer review happens for all outputs in an iterative manner, not just for the articles at the end.

In making and planning to share open results, you can apply the "as open as possible, as closed as necessary" principle. This means, protecting sensitive information, managing data protection practices where necessary and not carelessly sharing sensitive data or people's private information that can be misused. Online repositories, such as GitHub and GitLab, allow online interaction in addition to serving the technical purpose of version control and content hosting. For example, you can use issues and a project board to communicate what is happening in a project at any given point. The use of Pull Requests signals an invitation for peer review on the new development of code or other content. Thanks to a number of reusable templates you don't have to set up repositories from scratch. For example, you can directly use a template for reproducible research projects.

24.5 Role of Contributors in Open Science

Collaboration is central to all scientific research. The positive impact of collaboration is achieved when diverse contributors are supported to combine a range of skills, perspectives and resources together to work towards a shared goal. Projects that apply open and reproducible approaches, make it easier for diverse contributors to be involved and get recognized for their contributions while supporting the development of solutions that they can all benefit from.

Involving and recognizing the roles of all contributors in making open results is an important part of open science, which we will discuss next.

24.5.1 EHTC Case Study: Recognizing All Contributors

A map of the EHT. Stations active in 2017 and 2018 are shown with connecting lines and labeled in yellow, sites in commission are labeled in green, and legacy sites are labeled in red. From Paper II (Figure 1). IOPscience. https://iopscience.iop.org/journal/2041-8205/page/Focus_on_EHT

The Event Horizon Telescope (EHT) team involved 200 members from 59 institutes in 20 countries, from undergraduates to senior members of the field. They used an array that included eight radio telescopes at six geographic locations across the USA, Latin America, Europe and the South Pole. All collaborators were located in different geographic locations, had access to different instruments, collected data generated from telescopes in different locations and applied skills from across different teams to create groundbreaking results. Each contributor was acknowledged across different communication channels and given authorships in publications. EHTC also supports the "critical, independent analysis and interpretation" of their published results to facilitate transparency, rigor, and reproducibility (EHTC website).

24.5.2 Making Open Results Starts with Contributors!

Making different research components and preparing to share them as open results involve a range of activities. Behind these activities are the contributors who engage in various responsibilities that include, but are not limited to:

- Conceptualizing the idea
- Designing the project
- Serving as advisor or mentor
- Conducting experiment as a student, researcher, or research assistant
- Creating tools essential for carrying out the research
- Providing data expertise
- Developing software
- Providing specialized expertise and support
- Managing community and project requirements
- Providing feedback to the results
- Designing experiments and interpreting results
- Manuscript writing and review
- And more!

Too often conversations about contribution and authorship take place towards the end of a project or when a scientific publication is drafted. However, as you learned in the previous lessons, research outputs are generated throughout the lifetime of a research project. Therefore, it is important to build an agreement at the beginning of the project for how contributorship in the project will be managed.

Developing contribution guidelines and contributor agreements requires collaboratively defining what is considered contributions in your project, who among the current contributors will get authorship, who will get acknowledged as a contributor, what is the significance of the order in which authors are listed in a scientific publication, and who makes these decisions. Ensuring that all collaborators understand and agree to these guidelines before beginning the project is also important.

24.5.3 Contributors and Authorship

First and foremost, you must ensure that anyone who has contributed to the research project has their contributions recognized. With that shared understanding, in this lesson, you will explore what those recognitions as contributors or authors in your research project might look like.

Let's first define contributor and author roles.

A "CONTRIBUTOR"

AN "AUTHOR"

A contributor is anyone who has done any activity that made it possible for the research to happen and results to be created, published or shared.

A "CONTRIBUTOR"

AN "AUTHOR"

An author of an open result is a contributor who has given a substantial contribution to the conception or design of the work or the acquisition, analysis, or interpretation of the data for the published work.

24.5.4 Are All Authors Contributors and Vice Versa?

An author is a contributor who actively carries out one or several of the tasks listed above (National Institute of Health - NIH and ICMJE). All authors are contributors, but all contributors may not be authors, for example, someone serving as a mentor, trainer or infrastructure maintainer. Ideally, all contributors are given the opportunity to author research outputs.

Given the importance traditionally placed on authorship in scientific publication and the fuzziness of the definitions (that often contain relative terms such as "substantial" or "extensive" leaving too much room for interpretation), it is not surprising that determining who amongst the contributors gets to be an author can lead to biased or unfair decisions, disputes between contributors, or at the very least leave someone resentful and feeling unappreciated.

There is no single approach for recognizing contributors as authors, but here is what you should consider:

GROUP POWER DYNAMICS & EQUITY (E.G. SENIORITY, SYSTEMS OF OPPRESSION)

THE TYPE OF CONTRIBUTION

Consider this hypothetical scenario: You are a postdoctoral fellow and the leading author of a research project. A rotating student spends 4 months in the lab helping you set up and perfect the experimental protocol that you will then use to carry out the experiments needed

to answer your research question. They may even help you collect some preliminary data, but then they leave and later decide to join another lab. Would you provide authorship for the student?

It would be unethical not to give authorship or credit to someone who has provided significant help and contributed to the success of a research, even when they are no longer involved. A fair path in this scenario could be to contact the previous contributor and involve them in writing a relevant section of the manuscript.

GROUP POWER DYNAMICS & EQUITY (E.G. SENIORITY, SYSTEMS OF OPPRESSION)

THE TYPE OF CONTRIBUTION

The NIH guidelines for authorship outline what type of contribution does or does not warrant authorship. Each contribution is represented on a sliding scale and has no rigid cutoffs. Some contributions are given more weight than others. For example, for "design and interpretation of results", nearly all types of "original ideas, planning, and input" result in authorship. Whereas simply supervising the 1st author usually does not result in authorship (unless they are also contributing to the paper, of course). This is just one example. You will need to think about what this looks like for your own work!

Clear communication about roles and responsibilities early in the project, and guidelines for how credit will be determined, can help mitigate some of these issues.

24.5.5 Diverse Role of Contributors

It is important to set a reference for each research team/project about different kinds of responsibilities and opportunities available for different contributors and how each of them are acknowledged. CRediT Taxonomy represents roles typically played by contributors to research in creating scholarly output. Below, we provide a table with research roles that extends the CRediT taxonomy to include broader contributorship (Sharan, 2022). Using this as a starting point, open dialogue and discussion among team members can be facilitated to set a shared understanding and agreement about diverse roles of contributors including authorship of publications. The distinction between contribution types can help set clear expectations about responsibilities and how they can be recognized in a project.

Research Roles

Definition

Project Administration

Management and coordination responsibility for the research activity planning and execution

Funding Acquisition

Acquisition of the financial support for the project leading to the research and publications

Community Engagement

Connecting with project stakeholders, enabling collaboration, identifying resources, and managing contributors interactions

Equity, Diversity, Inclusion and Accessibility (EDIA)

Inclusive approaches to collaboration and research, involvement of diverse contributors, accessibility of resources, consideration of disability, neurodiversity and other considerations for equitable participation

Ethics Review

Ensure that if the research project needs to undergo an ethics review process

Communications and Engagement

Communications about the project and engagements with the stakeholders beyond the project and institution

Engagement with Experts and Policymakers

Pre-publication review, external advisory board meetings, regular reporting, post-publication reporting, and reaching out to the relevant policy makers actively

Recognition and Credit

Assessing incentives, creating a fair value system, fair recognition of all contributors

Project Design

Technical planning, expert recommendations, supervision or guidance, developing project roadmaps and milestones, tooling and template development

Conceptualization

Ideas; formulation or evolution of overarching research goals and aims

Methodology

Development or design of methodology; creation of models

Software

Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components

Validation

Verification, whether as a part of the activity or separate, of the overall replication/ reproducibility of results/experiments and other research outputs - generalizable

Investigation

Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection

Resources

Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools

Data Curation

Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later reuse (including licensing)

Writing - Original Draft

Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation)

Writing - Review & Editing

Preparation, creation and/or presentation of the published work by those from the research group, specifically critical review, commentary or revision – including pre-or post publication stages

Visualization

Preparation, creation and/or presentation of the published work, specifically visualization/ data presentation

Supervision

Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team

24.6 How to Give Open Recognition

To openly and fairly recognize all contributors, their names with the types of contributions they made should be listed in the project documentation. In manuscripts, it is a common practice to mention contributors' roles under the 'acknowledgement' section, such as using CRedIT or similar taxonomy as provided in the table above. All contributors should be encouraged to provide ORCiDs associated with their names to make them identifiable.

Contribution statements in documentation and manuscripts can specify who did what in the official results. This is great for transparency. It is also a great way to guard against unfair power dynamics. Details about contribution type shows explicitly who works on which parts

of results, and makes it easy to give fair authorship. For example: *"Pierro Asara: review and editing (equal). Kerys Jones: Conceptualization (lead); writing – original draft (lead); formal analysis (lead); writing – review and editing (equal). Elisha Roberto: Software (lead); writing – review and editing (equal). Hebei Wang: Methodology (lead); writing – review and editing (equal). Jinnie Wu: Conceptualization (supporting); Writing – original draft (supporting); Writing – review and editing (equal)."*

If a GitHub repository and website exist, a dedicated page should be created to list and recognize all contributors. If someone minorly contributed to the paper, code or data, you could add them as an author or contributor to the GitHub and Zenodo releases respectively. Engaged collaborators and contributors not already involved in making research outputs should be given the opportunity to contribute to open results such as through presentation, posters, talks, blogs, podcasts, data, software as well as articles.

24.6.1 Activity 3.1: Draft a Contribution Guideline

A standalone contribution guideline should be created for each open project, even when that means reusing an existing draft that the research team has used in another project.

Note that this is different from "contributing" guidelines that describe "how" to contribute (for example on code repositories). Contribution guidelines should describe contribution types and ways to acknowledge them as discussed above.

Contribution guidelines are not set-in-stone, but rather:

- Are discipline-dependent
- Can be adapted for your unique situation

You can begin by reviewing guidelines by NIH and ICMJEs for authorship contributions.

Notice that many categories and criteria for authorship, such as represented in the NIH guidelines' sliding scale, may be differently decided. For example, in some fields providing financial resources for a research project always warrants authorship. In other fields this is not the case.

Some projects may not follow traditional manuscripts as their outputs. For example, if software is a primary output from a project, there may be a need to define specific roles regarding code contributions. You can work with your research team to create a version of CRediT Taxonomy for your project, such as shared in an expanded version of the table above.

When different kinds of contributorship have been identified, clarify how different contributors will be involved and acknowledged. This may include recommended communication and collaboration processes for the team members, as well as recognition and credit for different kinds of contributions they make.

Additional Information

For additional tips on how to acknowledge different kinds of contributors to developing a resource including authorship, check out Acknowledging Contributors The Turing Way.

If working with online repositories such as GitHub, an app like 'all-contributors' bot is a great way to automate capturing all kinds of contributions, from fixing bugs to organizing events to improving accessibility in the project.

More systematic work is being undertaken by hidden REF who constructed a broad set of categories that can be used for celebrating everyone who contributes to the research.

There are several infrastructure roles like community managers, data stewards, product managers, ethicists and science communicators, who are also being recognized as valued members in research projects with an intention to provide leadership paths for technical and subject matter experts, even when their contributions can't always be assessed in tangible or traditional outputs [Mazumdar et al. 2015, Bennett et al., 2023].

The Declaration on Research Assessment (DORA) is also a good resource to understand what researchers, institutions, funders and publishers can do to improve the ways in which researchers and the outputs of scholarly research are evaluated.

24.7 Combining Open Results for Scientific Reporting and Publications

Scientific publications have traditionally remained one of the most popular modes of reporting and publication. Over the last decade, it has become a standard practice to submit pre-peer reviewed manuscripts on preprint servers (such as arXiv) to speed access to research before the peer-reviewed journal articles are published (discussed in Lesson 2). The publication system has also evolved massively. Journal articles are no longer about writing overview and summary of research, but can be used to share articles on software, data, education materials and more.

24.7.1 EHTC Case Study: Capturing Results on Activities Ranging From Collaboration to Observations, Image Generation to Interpretation

The polarized image of the M87 black hole shadow as observed on 2017 April 11 by the EHT (left panel) and an image from the EHT Model Library with a MAD magnetic configuration (right pane), with a list of papers describing different sets of results.

Across several preprints and eight peer-reviewed letters, EHTC presented open results issued from different teams on instrumentation, observation, algorithm, software, modeling, and data management, providing the full scope of the project and the conclusions drawn to date.

Open results such as reports, publications, code, white papers, press releases, blog posts, videos, TED talks and social media posts add to the comprehensive repertoire of open results supported by EHTC. Resources are centralized on the EHTC website, GitHub organization and YouTube channel among others to provide easy access to all open results.

It's important to highlight that their efforts have led to independent reanalysis and regeneration of black hole images. Specifically, Patel et. al. (2022) not only reproduced the original finding, but also contributed additional documentation, code, and a computational environment as open-source containerized software package to ensure future testing. Some of the original authors reviewed this work and made their comments also available online (Authorea).

24.7.2 How Do I Connect Open Results to Make Reproducible Publications

If not considered from the start, it can become challenging to ensure result reproducibility at the publication stage. Assuming that you have maintained open results considering their reproducibility, you can start assembling them to connect with the final reporting and publication with appropriate references to previous studies.

- Before writing your manuscript, assess each output to make sure that appropriate license is attached for reuse, documentation has been provided and contributors are clearly listed. You can decide to create a version of the record and point to a permanent identifier such as via Zenodo so that the link never breaks when sharing them on a public repository (such as GitLab/GitHub) or manuscripts with a visible list of contributors.
- Your publications can be created individually (such as in EHTC case study) or by combining several outputs or pieces of information in manuscripts. These will include resource requirements, dependencies, software, data, repository where code is shared with documentation and contributor information, among other research artifacts.
- The manuscript itself will describe research questions, methods as well as individual figures and tables explaining the results. When writing a manuscript, you can begin with figures by packaging data, code and parameters used, ensuring that information represented can be reproduced. You can find a detailed checklist in the publication by Gil et al. (2016).

As demonstrated in the EHTC case study, a final step towards making open results could be to create a meta article and/or simple website/git page that centralizes all your research outputs. Different parts of research (individual open results) can be accessed centrally with details including open recognition for all contributors.

If you are looking for concrete actions you can take to make open results, pick one of these four items:

- Improve how you define contributorship in your project and how authorship is assigned.
- Ensure the data or software in your paper is uploaded to Zenodo with license and documentation including metadata and that the DOI is posted to your scientific report and publication.
- Ensure that the process you use to collect data and perform its analysis, including all the dependencies and methods used in your data analysis pipeline, are clearly described to allow others to reproduce your results.
- Create a centralized repository or a simple git page to centralize all research outputs with contributors list.

24.8 Lesson 3: Summary

The steps that we highlight to make open results are not intractable. In fact, the steps we have highlighted are things we can do on a regular basis to ensure that all research artifacts can be shared later as open and reproducible results. In this lesson we learned:

- Approaches for making open results.
- The importance of collaboration in making results.
- How to recognize and credit all of the contributors who make results.
- How to combine different open results to create scientific reports and reproducible outputs.

24.9 Lesson 3: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/02

1. Which of the following roles would be most appropriately credited with contributorship? Select all that apply.

- Original idea, planning, and input
- Supervision of the project
- Original experimental work
- Data analysis
- Drafting of manuscript

Question

02/02

What is not an example of open research results?

- Open access papers
- Conference presentation
- Internal team meeting notes
- Regular reports shared online
- Poster at a workshop
- Blog post
- Computational notebook on GitHub
- Figure with a DOI (e.g, Zenodo or Figshare)
- Pre-print of a paper

25 Lesson 4: Sharing Open Results

25.1 Navigation

25.2 Overview

In lesson 3 you learned about how to make reproducible results. Now, we can finally think about how to best share those results. In this lesson we will place emphasis on publishing manuscripts as open access. You will learn what subtleties to consider when determining what journal to publish in, including how to make sense of a journal's policies on self-archiving. Finally, we discuss some commonly held concerns about sharing open access publications, and how to overcome them. Ultimately, we want to ensure that you have confidence in your decision to publish as open access.

25.3 Learning Objectives

After completing this lesson, you should be able to:

- List ways that you can share open results to become a more collaborative, effective, scientist.
- List different types of open access publications and considerations when sharing like licenses.
- List some of the concerns around open access publishing, including responsibilities for authors, the threat of predatory publishers, and the fear of being wrong.

25.4 When to Share

Part of doing open science is enabling collaborative, interactive results. Sharing different types of research objects earlier in your research process helps increase visibility to your work and accelerates your efforts by drawing from the collective knowledge of others. The internet has fundamentally changed the timing of and manner in which scientists communicate results.

Planning to share your intermediate results at the beginning of your project makes sharing final results easier. The figure above illustrates many of the different objects that can be shared before the 'final' report or publication. Sharing and talking about your research as you are doing it, as well as engaging with other scientists, will increase the robustness of your work.

Ask questions. Share what you are working on. You will find that many involved in the scientific community want to help. The more you engage, the larger the audience and the more impact you will have when that 'final' publication is published.

In the past few decades, scientists have made new connections and sought collaborators through letters and at conferences. However, this way of doing science tended to restrict who could participate. Today, most of these discussions take place on the internet, which has enabled new avenues for participatory science, open to all.

The platforms where you share research depends on what you want to share. Reference the figure above and think about where you might share different types of information. How will this influence who you have an ability to engage with?

Let's start with sharing in smaller groups (workshops and conferences) and move to larger audiences. There are distinct reasons for communicating results to different sizes of groups, as explored in the following sections.

25.4.1 At Workshops and Conferences

Many of us attend scientific conferences, workshops, and other gatherings to discuss our science with peers. The costs associated with attendance and travel to these events may limit who has access to the material presented there. At these events, scientists often give talks or present posters that are not yet peer reviewed to invite feedback from the community and potentially recruit collaborators. These interactions are important for improving research projects, and are often done when a project is still ongoing so that researchers can gather feedback early in their scientific process.

It is important to think about what audience you will be reaching at an event. Conferences have different policies about open access to materials presented at an event. Consider what you are sharing and who you want to share it with. For example, not all events provide long-term open access to workshop materials after the event. If you want to reach a larger audience or preserve the materials long-term, as a scientist, you have options to license and publish

presented materials yourself (for example using Zenodo with a DOI) if an event doesn't do so.

25.4.2 Other Forms of Interactive Feedback

Other forms of sharing can serve a similar purpose to share and document your results and/or software packages, and also allow for additional flexibility and openness! There are a number of additional resources that you can use

- Blog posts and online articles
- Short form videos and podcasts
- Computational notebooks
- Social media posts
- Forum discussions

These different pathways allow for the dissemination of null results, intermediate science updates and/or software improvements. These alternative ways of sharing your work can benefit your research by facilitating extended dialogue between you and collaborators, and even the general public. Additionally, the public has easier access to these forms than they do to conferences.

Here are some specific examples of engagement across contemporary platforms for scientific collaboration:

- Blog posts such as the Pangeo blog - see examples of how to use different software tools for different science questions!
- Computational notebooks as a way to demo software techniques (e.g. the Project Pythia Cookbook Gallery showcasing computational science workflows in the Earth sciences).
- Non-peer reviewed publications, such as Research Notes of the AAS.
- Team and/or Mission Science Pages, such as the LUVOIR team's page or the Juno mission's page.
- Conference proceedings, such as from the Society of Photo-Optical Instrumentation Engineers.
- Social media posts: https://twitter.com/MartianColonist/status/1706824699349488036

Over the course of a 3 year study, the Journal of Medical Internet Research found that highly tweeted articles were 11 times more likely to be highly cited than less tweeted articles.

25.4.3 Publishing Reproducible Reports and Publications

An open access report and paper can be reproducible when its data, software, and content are made available to the readers following best practices. There is a growing list of resources documenting how to make open results reproducible (such as The Turing Way and FORRT).

There are several examples (discussed in these lessons) that demonstrate how we can integrate technical and collaborative solutions to enable reproducibility. For example, executable notebooks allow interactivity and testing, training workshops invite feedback for improvement and GitHub/GitLab enable community based open review.

Scholarly Journals

Publishing work in a peer-reviewed journal forms the traditionally written basis of how we share our science, and is important for communicating scientific detail and rigor to colleagues. Academic journals also act as a long-term archive of scientific research papers. For many scientists, publishing in peer-reviewed journals and receiving citations are key factors in how they are evaluated for career advancement, positions appointments, committee memberships, and honors.

Traditionally, authors pay an Article Processing Charge (APC) that can range from $200-$12000 USD. Higher profile journals often charge higher fees to authors. Accessing articles has traditionally been restricted by pay-walls that require a subscription or charge per article. Journals have different options for making your published work accessible to various communities.

Who Has Access to Journal Subscriptions?

Paywalls limit who can access scientific research. This barrier acts to limit who can participate in science and erodes public trust in results. Part of open science is ensuring worldwide access to research.

Open Access Journals

Open access journals are peer-reviewed journals that are more accessible because they don't require readers to have a subscription or pay to access the content. However, open access journals often require additional fees for the author. Open access peer-reviewed articles are archived by a more formal discussion of scientific ideas, interpretations, and conclusions. They form the basis of how researchers share results.

25.4.4 Activity 4.1: Read the Open Access Policies of Publishers That You Use

In this activity, you will learn how to access information about a journal's data archive policies. The Directory of Open Access Journals (DOAJ) provides an extensive index of open access journals around the globe. The DOAJ can be used to look up information, including data

archiving policies, for journals that publish research. Let's open up this website and look up the policies specific to your most-used journals.

1. First, navigate to the DOAJ website.
2. Type in the name of one of the following journals in the search box, and then click on the yellow "SEARCH" button.

- Atmospheric and Oceanic Science Letters
- Swiss Journal of Geosciences
- History of Geo-and Space Sciences Note: You may input any journal desired but for this exercise use one of those listed to see the Sherpa/Romeo link that is listed in Step 5.

3. The search results may show more than one match. Select the desired journal within the search results by clicking on the journal name.A dashboard appears, giving information regarding publication fees, waiver policies, the type of open license used, and other information on multiple displayed titles.
4. Click on the "archiving policy" link appearing in one of the displayed boxes as seen here. This will provide links to extensive information regarding the journal's open access policies for the manuscript itself: An extensive amount of information will be presented, including details on the publishing policies specific to the selected journal.
5. Alternatively, to get a more condensed view of the journal's policies, return to the DOAJ dashboard on the About page with the multiple boxes displayed, and click on the "Sherpa/Romeo" link as shown here.
6. On the Sherpa Romeo page, click on the journal name that is displayed in the list (the only journal displayed).
7. When you view the page, you see that it consolidates and summarizes the open access policies for that journal and associated materials. The published version is likely to be the most relevant (see red box in figure).
8. Review the page and determine which license the journal you selected has defined for reusability for manuscripts.

25.4.4.1 Activity Key Takeaways: Read the Open Access Policies of Publishers That You Use

This is an example of a site that you can use to determine if a journal's policy is consistent with how you wish to publish your open access results. Journal policies should always be reviewed and considered during the early planning phase of your project and well before submitting your manuscript for publication.

25.5 How to Share

Perhaps the single most important step to make your results open is to assign them a globally unique and persistent identifier. This will give you a single code, URL, or number that you can use to uniquely refer to a research object. Any derived research object can use this identifier to link to it and create a traceable and rich history of use and development. Crucially, this identifier can be used by others to cite and credit your work (source).

The identifier must also be persistent. This guarantees that the identifier points to the same research object for a long period of time. What counts as "persistent" is, of course, a matter of degree since even the most stable identifier probably won't survive the Sun engulfing the Earth in a few billion years. In this context, "persistent" implies that it is registered in a database managed by an organization or system that is committed to maintaining it as stable and backwards compatible for the foreseeable future.

For example, URLs (for example, a personal website, GitHub repository, or cloud storage) are notoriously not persistent since they can change their contents frequently or become invalid without maintenance. On the other hand, Journal publications have a Digital Object Identifier (DOI) whose persistence is guaranteed by the International DOI Foundation.

As well as uniquely identifying each research object, it is important to be able to uniquely identify and cite all the authors and contributors. For this, it is recommended to get the permanent digital ID of each of the authors and contributors. ORCID (Open Researcher and Contributor ID) is an online service where you can get a permanent digital identifier.

There are examples of globally unique and persistent identifiers:

DIGITAL OBJECT IDENTIFIER10.1371/JOURNAL.PONE.0230416

ISBN-13: 978-0735619678

THE INTERNET ARCHIVE

The Digital Object Identifier is provided by the International DOI Foundation, which ensures that each ID is unique and ensures that a DOI link always links to the correct object.

DIGITAL OBJECT IDENTIFIER10.1371/JOURNAL.PONE.0230416

ISBN-13: 978-0735619678

THE INTERNET ARCHIVE

This is an International Standard Book Number, which has to be purchased by publishers by the International ISBN Agency.

DIGITAL OBJECT IDENTIFIER10.1371/JOURNAL.PONE.0230416

ISBN-13: 978-0735619678

THE INTERNET ARCHIVE

The Internet Archive captures snapshots of websites and their links are really stable. Even if not ideal, it's a handy tool for creating identifiers of websites easily.

25.5.1 Licenses

By applying a license to your work, you make clear what others can do with the things you're sharing, as well as the conditions under which you're providing them (like the requirement to cite you). Another very important element to include with your research objects is clear rules for reuse (as is and for creating derivative work), which are often and most easily codified by the use of licenses.

Without a license, all rights are with the author of the research result. That means nobody else can use, copy, distribute, or modify the work without consent. A license gives this consent. If you do not have a license for each of the research objects that constitute your research result, it is effectively unusable by the whole research community.

Creative Commons licenses are usually used for written content (see Lesson 3 for a full description!). The benefit of a license, as opposed to the public domain, is that most require attribution to the original creators. The Creative Commons Attribution License, CC-BY, is the most common open access license for sharing publications as it requires attribution. There are other Creative Commons licenses used that may have different limitations on whether or not they can be commercially used, whether or not they can be modified and copied, and whether or not the licenses can be changed in further adaptations of code.

Your institutions, funding agency, or research proposal may require use of a specific license depending on the type of material that you produce from your research. For public agencies, CC-0 or CC-BY are generally recommended (or required) to maximize their return on investment and ensure widest possible re-use. Choosing a CC license that has additional restrictions (eg. -ND, -SA, -NC) can result in less reuse of data. As you share results on different platforms, look carefully to see what license is being applied!

25.5.2 Routes for Open Access Publishing

Routes to publishing openly. The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 license. Original version on Zenodo. http://doi.org/10.5281/zenodo.5706310

The most common types of open access publishing are Green, Gold, and Diamond.

GOLD OPEN ACCESS PUBLISHING

GREEN OPEN ACCESS PUBLISHING

DIAMOND OPEN ACCESS PUBLISHING

In Gold Open Access Publishing, authors pay an Article Processing Charge (APC) to a journal so that they publish the final version of your article under an open access license, which is then permanently and freely available online for anyone. The author will retain the copyright of their article, usually via a Creative Commons license of their choice, which dictates what others can do with the article. A criticism around gold Open Access publishing is the cost.

APCs can generally be around 2000 USD or in some cases more, which can therefore be prohibitive for authors across the globe. Some publishers offer discounts or waivers to authors from countries classified by the World Bank as low-income economies or APCs may be covered by your funder as part of your grant.

GOLD OPEN ACCESS PUBLISHING

GREEN OPEN ACCESS PUBLISHING

DIAMOND OPEN ACCESS PUBLISHING

Green Open Access is the process of self-archiving. The self-archiving movement aims to provide tools and assistance to scholars to deposit and disseminate their refereed journal articles in open institutional or subject-based repositories. You may choose to self-archive your work to make it more discoverable and/or after you've published it in a subscription journal to ensure there is an open version of your paper.

The Registry of Open Access Repositories contains a list of repositories that are available for researchers to self-archive. At the beginning of 2019, there were more than 4000 repositories. It is important to find yourself-archive community!

GOLD OPEN ACCESS PUBLISHING

GREEN OPEN ACCESS PUBLISHING

DIAMOND OPEN ACCESS PUBLISHING

Diamond Open Access are publications where there is neither a cost for reading the article or publishing an article. Diamond Open Access journals either have very low costs due to building on existing infrastructure and volunteer efforts, or are supported directly by foundations or institutions. For authors, Diamond Open Access publications typically allow the author to retain copyright and the final version of their article as it is published under an open access license.

25.5.3 Pros and Cons of Preprints

When publishing in a peer-reviewed journal, you can decide to share a pre-print. A preprint is a version of a paper prior to its publication in a journal*. This can be the author's version of the accepted manuscript after peer review or a version prior to submission to a journal.

The accepted manuscript is the final, peer-reviewed version of the article that has been accepted for publication by a publisher. The accepted manuscript includes all changes made during the peer review process and contains the same content as the final published article, but it does not include the publisher's copyediting, stylistic, or formatting edits that will appear in the final journal publication (i.e., the version of record).

Many journals provide preprint services. If they don't, there are many public preprint servers available. Often the funding agency will have a preferred public preprint server.

Preprints come with many advantages as well as perceived or potential disadvantages.

ADVANTAGES TO PUBLISHING WORK AS A PRE-PRINT

POTENTIAL DISADVANTAGES

Quickly disseminate findings to communities in a timely manner.

Many field-specific preprint servers (e.g. arxiv.org, biorxiv.org, essoar.org) are free to both upload and read.

Community feedback on your work as it's being done.

ADVANTAGES TO PUBLISHING WORK AS A PRE-PRINT

POTENTIAL DISADVANTAGES

Work may be shared with critical errors that may have been caught in peer review.

In some fields, there is a perception of lessened reliability or quality of research published as a peer print.

Some journals do not allow or accept articles if they have been submitted to a preprint server.

25.5.4 What to Consider When Making Preprints

When deciding to preprint your work, you will need to check:

1. The copyright policy of the journal with which you aim to publish.
2. The version of the paper that can be deposited.
3. When the paper is allowed to be made publicly available.

25.5.4.1 Additional Reading:

Read the story about how Joanne Cohn's email list for preprints led to Paul Ginsparg's development of arXiv.

25.6 Other Considerations When Sharing

25.6.1 Who is Sharing?

When writing an OSDMP, it's important to include a plan for the roles and responsibilities needed to share your results. As discussed in lesson 3, your community will consist of members in different roles – some actively engaged, some with only a passing interest. Having a clear plan for sharing open results and how credit will be given will help everyone understand their contributions and roles and minimize conflict.

Lesson 3 describes in detail the different roles that people may play in sharing results. This should be clearly described In the OSDMP.

25.6.2 Predatory Publishers

Predatory Publishers are generally for-profit publishers that charge a publishing fee but provide few quality checks on the quality of the publication that would be expected from scholarly publications. They sometimes use the benefit of open access to entice authors to publish with them. If you are unsure if a publisher may be predatory, checking with your library staff is a good place to start.

There are many red flags in these requests for predatory publishers:

- There is an urgency and request for an extremely quick turnaround. A very fast publication time might indicate a less rigorous peer-review process.
- Written English in correspondence is often poor quality with many grammatical errors. (Though it's important to remember that this alone does not indicate predatory behavior, as grammatical mistakes can be made for innocent reasons, such as being a non-native speaker.)
- The journal subject is nonspecific.
- The solicitation is inaccurate or generic.
- The email is often unsolicited, even if they claim that they're referring to a previous paper of yours. This might start with an inaccurate or generic solicitation such as "professor".
- They emphasize ISSN indexing and/or impact factors, although this particular journal doesn't have one. Consider Journal Citation Indicator (JCI) in addition to Journal Impact Factor (JIF).

- The publisher/journal sends multiple emails soliciting manuscripts, special issues, and editorial roles.
- They have a high number of special issues, such that the majority of the papers published appear in special issues.
- Their name resembles the name of a prestigious journal.
- They have a high self-citation rate, such as over 20%.
- They have a very high acceptance rate of submitted papers.
- They send frequent requests to submit/serve as editor.

Below are some final thoughts on what or what not to consider when deciding where to publish. As with many considerations you will encounter in academia, sometimes deciding the best place to publish will be determined by word of mouth conversations with peers. Read more on NOAA's guidance on predatory publishing.

25.6.3 Common Questions About Sharing Results

Sharing in different ways, especially without peer-review, can be intimidating. Maybe you have worried about the following questions:

- **What if an open result is wrong?**A tweet, post, or video is only a snapshot in time of a research result. It is understood by all working professional scientists that we are constantly learning and discovering new things. Making reproducible results will necessarily include different versions and revisions of an idea as it develops.
- **I have already published my science as an open result, so do I need to respond to community feedback forever?**As long as you have done everything to make your work reproducible - you don't need to worry. Open science can't be carried solely by a single person. Open science communities can continue to update, refine, and develop your open science result if your work has been shared and openly licensed.If you are able to address a question or a concern about your prior research, that's great. It is also an ethical response to acknowledge that this is research that you are no longer actively involved with, but allow others to continue the work that you began.
- **What if I can't do everything? Am I a bad open scientist?**The short answer is no! You have only a limited amount of time. Even with collaborators, you can't possibly do everything.

Sharing open results improves science - it is faster, more accessible, and more collaborative. In this lesson you have learned about all the different ways you can share open results. Think about how you might share something you are working on now!

25.7 Lesson 4: Summary

In this lesson, you learned:

- When to share open results and the different ways in which they can be shared. This includes: peer reviewed publications, conference proceedings, blog posts, videos, notebooks, and social media.
- How to share open results including considerations around the license for the publication, routes for open access publications (Green, Gold, Diamond), and preprints as part of the publication process.
- Considerations around sharing, including considerations around predatory publishers and common concerns around openly sharing of results.

25.8 Lesson 4: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/03

Which of the following Creative Commons licenses is most commonly used for open access publications?

- CC BY-NC-SA
- Copyright
- CC-BY
- Apache 2.0

Question

02/03

Read the statement below and decide whether it's true or false.

Diamond open access is both free to publish and to read scientific articles.

- True
- False

Question

03/03

Take a close look at the request for journal submission below. Does this request for journal submission seem reliable?

- Yes
- No

26 Lesson 5: From Theory to Practice

26.1 Navigation

26.2 Overview

In the previous lessons, we learned about various ways to share our science, and what steps we should think about when sharing. In this lesson, we tie the concepts from previous lessons together with some specific guidance for writing the Sharing Results section of an Open Science and Data Management Plans (OSDMP). We will also reflect on how our society and technology constantly evolve, as does the way we do science. A new technology with the potential to radically alter the way we do and share science is artificial intelligence (AI), particularly when it comes to language learning models. These AI tools are already changing how we interact with written text. In this lesson, we discuss some of the ways that AI is and will affect how we do and share our science.

26.3 Learning Objectives

After completing this lesson, you should be able to:

- List what to include in an OSDMP for sharing results openly.
- List some concrete steps toward sharing results openly.
- Describe how emerging technology like AI is currently impacting how we use, make, and share our science.

26.4 Writing an OSDMP: What to Include in the OSDMP for Sharing Results Openly

The process within an Open Science and Data Management Plans (OSDMP) to share data and software is covered in other modules, so here we will discuss how to share the other type of research outputs. Most proposals require that you include plans for publications such as peer reviewed manuscripts, technical reports, books, and conference materials.

Though not required, it can be a good idea to include plans for making your results publicly accessible in ways other than traditional publishing, e.g. online blog posts, tutorials, or other materials. After all, writing an OSDMP is often required for funding requests, and this can be a way to show proposal reviewers that you are thinking about how to best share your science.

26.4.1 Activity 5.1: Pen to Paper

Write a sample results section of an OSDMP that details how you would plan to make your results open. Think about an example from your research and what details you would need to include to convince reviewers that you will share open access results.

Example 1: This activity will result in 2 peer-reviewed publications that will be published green open- access. Pre-prints will be archived in PubSpace.

Example 2: This activity will result in the creation of computational notebooks, 4 conference abstracts and posters, 2 peer reviewed manuscripts, and 2 online plain-language articles, summarizing our results. Peer-reviewed publications will be published green open-access and pre-prints will be archived in PubSpace or the journals open-access preprint server. All other materials will be archived at Zenodo, assigned a DOI, and assigned a CC-BY license or permissive software license.

For these examples, what other information or details could be added? If you were planning to write a tutorial about your science, what would you include?

26.5 Example Steps Toward More Open Results

NASA Announces Summer 2023 Hottest on Record

Image credit: NASA Earth Observatory/Lauren Dauphin.

When results and research objects are published openly, anyone can reproduce the scientific result. For topics like climate change, the transparency of results helps reduce misinformation and increases public trust in results.

Here is a GitHub repository with an example of a result made available as open access. This visualization is not perfect but provides a snapshot of a work in progress that can be shared with the community for feedback and refinement. This could be further refined, or perhaps serve as the start of a new effort that will extend the initial results. The results are more accessible, inclusive, and reproducible by being published openly.

There are lots of ways that open science can extend the span or scope of projects. Here are some steps you can take to share your open results in a way that makes your work more usable, reproducible, and inclusive:

- Add a Code of Conduct via the CODE_OF_CONDUCT file and link to other policies that apply to your work.
- Add contributors and authorship guidelines via a CONTRIBUTING file.
- Add your collaborators and team members' names with their permission.
- Add your proposal but remove any sensitive information.
- Create a preliminary roadmap and what goals the project is trying to achieve.
- Create a project management, code and data folders where you can upload appropriate information as your project develops.
- Create a resource list that your project requires.
- Provide links to training materials that your collaborators and contributors may benefit from.
- Use issues and project boards to communicate what is happening in the project.
- Use Pull Request to invite reviews to new development of code and content.
- Add user manual and executable notebooks to allow code testing.
- Create and share executable notebooks that document how data is processed and the result obtained.
- Create tutorials or short form videos demonstrating how a step in your research workflow was accomplished.
- Write a blog post about your experience wrestling with a particular research challenge and how you solved it.
- Contribute to documentation to improve the open-source tools based on your own experience.
- Connect your repository to Binder to allow online testing of your code and executable notebooks.
- Link all the outputs that are generated outside this repository (like blog, video, forum post and podcast among others as discussed above).
- Some advanced steps that should be applied as the project develops include continuous integration, containerisation, Citation CFF file and the creation of a simple web page to link all information.

26.6 How Emerging Technology Like AI is Changing How We Do Science

Throughout these modules, the internet has been identified as a fundamental disruptive technology that changed how almost all of science is accomplished. Scientists rarely go to libraries to read the latest journal articles. Data is no longer mailed around the world on tape drives. Software isn't shared via floppy disks. The internet helped create the modern scientific workflow and made science more interactive and accessible. Now AI tools are starting to disrupt science in a similar manner. AI is not only revolutionizing many aspects of our lives, it is also changing how we do science. As companies race to create and integrate new generative AI tools into every aspect of our lives, many scientists, institutions, journal publishers, and agencies are looking to see how to use these tools effectively, understand their reliability, accuracy, biases, and how to also use these cutting edge tools ethically. An additional concern is how any information shared with AI tools may be used to intentionally or unintentionally disclose confidential data, leading to privacy concerns.

AI can help us use and share research. It can act as an accelerant, taking care of tedious tasks while leaving scientits free for more creative thought. These tools are better than humans at processing vast amounts of data, but humans are better at creative and nuanced thought. This is important to consider when determining whether or not to use AI. As an example, many people already use AI tools to help with their inbox management and writing emails with AI generated suggested content. Within science, there are many potential tasks that could potentially be expedited using AI, according to three studies published in Nature:

- AI science search engines are exploding in number — are they any good?
- How AI technology can tame the scientific literature
- AI and science: what 1,600 researchers think

26.6.1 Using AI:

LITERATURE REVIEWS

SEARCHING FOR RELEVANT DATASETS AND SOFTWARE TOOLS

LANGUAGE BARRIERS

The ever-increasing volume of scientific literature has made it challenging for researchers to stay abreast of recent articles and find relevant older ones. AI tools can be used to create personalized recommendations for relevant articles as well as create summaries of them in various formats. Some examples of these tools include SciSummary, SummarizeBot, Scholarcy, Paper Digest, Lynx AI, TLDR This.

Possible drawbacks when using these tools include:

Potential introduction of biases

Insufficient contextual understanding or interpretation

Possible inability to handle complex technical language

Incorrectly identifying key points

LITERATURE REVIEWS

SEARCHING FOR RELEVANT DATASETS AND SOFTWARE TOOLS

LANGUAGE BARRIERS

AI tools can be used to discover different datasets that may be relevant to a scientific query and recommend relevant software libraries.

LITERATURE REVIEWS

SEARCHING FOR RELEVANT DATASETS AND SOFTWARE TOOLS

LANGUAGE BARRIERS

AI tools can be used to create automatic translations into different languages. Several of the tools above also offer translation.

26.6.2 Making with AI:

CODE

RESULTS

AI tools can be used to generate code to perform analysis tasks and translate between programming languages. Some examples of these tools include Co-Pilot, Codex, ChatGPT, and AlphaCode.

Usage tip: Popular large language models can be used to generate code, but it has been noted by many that breaking down tasks and using careful prompts helps generate better results.

CODE

RESULTS

AI tools can be used to generate text, summarize background materials, develop key points, develop images and figures, and conclusions. Using these tools may help non-native speakers communicate science in different languages more clearly. Additionally, they could be helpful to develop plain-language summaries, blog posts, and social media posts.

Some possible drawbacks when using these tools:

See the list above for a literature review.

Factual and commonsense reasoning mistakes because they do not (at this time) have the type of cognition or perception needed to understand language and its relationship to the external physical, biological, and social world (cite: https://www.tandfonline.com/doi/full/10.1080/08989621.2023.21685

26.6.3 Sharing with AI:

- Results - AI/ML models are increasingly being used in research. When sharing results, follow best practices as outlined in the Ethical and Responsible Use of AI/ML in the Earth, Space, and Environmental Sciences article.
- Incremental prompting can help create an outline for your research article. An example can be found on X.
- AI tools can help identify where to share results and help write social media or other short posts based on your article.

26.6.4 Cautions About Use of AI Tools

Journals are increasingly implementing guidelines and requirements concerning the usage of AI tools during the writing process. Many require that the use of AI tools for writing, images creation, or other elements must be disclosed and their method of use identified. As is the case with all other material within an article, authors are fully responsible for ensuring that content is correct. Examples of this policy can be read in the AI guidelines of Nature and NCBI.

Furthermore, there are numerous examples of generative AI (for both code and content) delivering plagiarized information in violation of licenses, as well as fabricateding material including citations. Using these AI tools may lead to findings of academic and research misconduct should fabrication, falsification or plagiarism be contained within AI generated materials. So BE CAREFUL. Learn more about possible issues with AI in a Nature example here.

At this time, and for these reasons, AI tools are generally not allowed in grant applications or in peer- review or proposal review activities.

The National Institutes of Health (NIH) has prohibited "scientific peer reviewers from using natural language processors, large language models, or other generative Artificial Intelligence (AI) technologies for analyzing and formulating peer review critiques for grant applications and R&D contract proposals." Utilizing AI in the peer review process is a breach of confidentiality because these tools "have no guarantee of where data is being sent, saved, viewed or used in the future." Using AI tools to help draft a critique or to assist with improving the grammar and syntax of a critique draft are both considered breaches of confidentiality. Read NIH's AI policy here.

AI tools for science are developing rapidly. The science community's understanding of how to ethically and safely use AI is just developing as its use in research expands rapidly. The

guidelines above offer a snapshot in time and will likely continue to evolve. If you choose to use these tools for scientific research, carefully consider how much to rely on them and how their biases may impact results, as cautioned in this Nature article. The internet has transformed the world and AI tools are likely to do the same. As with any tool, it is important they are used for the appropriate purpose and in an ethical manner.

26.7 Lesson 5: Summary

The steps that we highlight to make your research more reproducible and open will advance science and the impact of your research. In fact, the steps we have highlighted are things we can do immediately to ensure we make open and reproducible results.

In this lesson, you learned:

- How to include open results in the OSDMP.
- An example of how results can be shared openly.
- That developing AI tools are being used in all parts of the scientific workflow, they are changing rapidly, and there are still many open questions about how and when to use them.

26.8 Lesson 5: Knowledge Check

Answer the following questions to test what you have learned so far.

Question

01/03

Read the statement below and decide whether it's true or false.

It is a good idea to include plans in your OSDMP for making your results available in ways outside of traditional publishing, e.g. online blog posts or tutorials.

- True
- False

Question

02/03

Which of the following aspects of AI are considered as benefits? Select all that apply.

- Personalized journal article recommendations based on your discipline and interests
- Recommendations for data and software relevant to your science project
- Potential introduction of bias

- Factual mistakes
- Translation between languages

Question

03/03

Which of the following are steps you can take to share your open results online? Let's assume that, like the activity, you are sharing an interactive visualization.

- Host your project in a public GitHub repository
- Assign an open license
- Add a code of conduct to the GitHub repository
- Add a user manual
- Release your project on public repositories that assign DOIs
- All of the above

26.9 Open Results Summary

26.9.1 Moving Toward an Open, Collaborative, and Inclusive Scientific Future

Science is meant to benefit society. Sharing our science helps ensure that it benefits society and informs the decisions of the public and policymakers, especially when funded by public agencies or governments. Going back to the 'Ethos of Open Science' module:

"Open Science is the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility, and equity"

https://open.science.gov/

Throughout this curriculum, we have focused on skills needed to make research products and processes available to all. The traditional practice of only sharing results limits insight into how science is done and may act to limit who can participate in science. By sharing your scientific process and working openly, you advance all of science in a more rapid and inclusive way. This curriculum will continue to evolve as science evolves and we welcome your contributions!

26.10 Open Science 101 Summary

Congratulations! You have successfully completed Open Science 101! Thank you for taking the time to learn about open science - you are part of a broader movement to improve science and make our world better!

Ready to learn more? Here are some great next steps:

26.10.1 Learn more about and engage with TOPS!

TOPS website

CLICK TO LEARN

TOPS GitHub Discussion Forum

CLICK TO LEARN

26.10.2 Learn more through online courses:

OpenSciency

CLICK TO LEARN

Open Science MOOC

CLICK TO LEARN

26.10.3 Take your coding and data science skills to the next level!

Carpentries

CLICK TO LEARN

26.10.4 Read online guides and learn about ongoing open science community initiatives:

The Turing Way

CLICK TO LEARN

Center for Open Science

CLICK TO LEARN

Open Science NL

CLICK TO LEARN

These are just a start - there are a lot more fantastic open science resources online!

Part VII

About Transform to Open Science

What We Do

The world is changing rapidly. Everyday new problems emerge and it takes groundbreaking scientific discoveries to solve them. To stay ahead, the pace of science must accelerate and science needs to be even better, more accurate, and faster to enable the truly transformative breakthroughs that will help us thrive. Closed science, hoarding information and resources, silos of knowledge holds science back by limiting who can participate. We need more voices that work together and share knowledge and resources. Only then will we find new and better solutions. This Transform to Open Science (TOPS) mission will allow us to create a scientific culture that is ready for 21st century challenges. Open Science will **broaden participation, increase accessibility to knowledge, and embrace new technologies** that can respond to these changes at scale. We hope you will join us in creating an open science infrastructure in your organization.

Strategic Objectives

Open Science creates more advanced and inclusive research faster, builds a more just and equitable world, and ensures that minds from all walks of life can participate in science. TOPS is an ambitious plan to accelerate open science practices. It's a 5-year journey that will
1. Accelerate major scientific discoveries 2. Broaden participation by historically excluded communities 3. Increase understanding and adoption of open science principles and techniques

We hope you will join us and champion open science!

27 Frequently Asked Questions

This page is a living document of the most common questions posed to the TOPS team. It is our hope that these answers will benefit the wider community.

What is open science?

The federal government defines open science as the principle and practice of making research products and processes available to all, while respecting diverse cultures, maintaining security and privacy, and fostering collaborations, reproducibility and equity.

What is the difference between open-source science and open science?

The primary difference is that open-source science commits to making the scientific process open from the start of research activities rather than making research results open once the research is complete and papers are published. The commitment to conduct research in the open supports greater participation in answering fundamental scientific questions and the use of publicly funded research, data, and analysis for societal benefit.

In what ways has open science contributed to accelerating the pace of scientific advancements?

By removing barriers to access, such as paywalls and restrictive licensing, researchers worldwide can freely access a wealth of information, fostering interdisciplinary collaborations and the exchange of diverse perspectives. This facilitates faster dissemination of findings, enabling scientists to build upon existing research and push the boundaries of innovation more rapidly.

What makes open science so successful?

Open science's success lies in its ability to create a more inclusive, transparent, and collaborative research environment, driving scientific advancements and societal impact. By allowing for broader sharing of research data and methodologies worldwide, and encouraging collaboration, open science can accelerate scientific progress while enhancing credibility and reproducibility.

How has open science improved transparency and trust in the scientific research process?

One of the fundamental pillars of open science is transparency, significantly improving trust in the scientific research process. Through open access to data, methodologies, and research

outcomes, the scientific community fosters greater accountability and credibility. This transparency cultivates trust among peers, institutions, policymakers, and the public, enhancing the integrity of scientific endeavors.

What benefits have researchers experienced by openly sharing their data and findings with the broader scientific community?

Openly sharing research data and findings accelerates scientific progress by fostering collaboration, increasing visibility and impact, and promoting transparency and reproducibility. This practice leads to faster advancements, innovative discoveries, and broader societal impact while complying with funding policies and enhancing the credibility of research outcomes.

In what ways has open science played a role in enhancing the reproducibility and reliability of scientific studies?

Open science has played a pivotal role in enhancing the reproducibility and reliability of scientific studies. Transparent methodologies, open data, and shared protocols enable independent validation and replication of research findings. This fosters a culture of rigor and robustness, ensuring that scientific conclusions are built on sound evidence and are more trustworthy.