AGAPE: An introductory course to open science for early career researchers
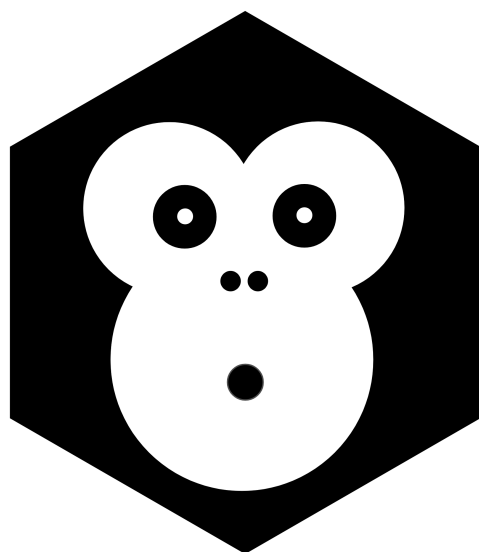
An Agape initiative

Contents

Introduction

Greetings, fellow early career researchers and open science-curious friends!

In this course, we will introduce you to the world of open science. Perhaps you are familiar with some of the concepts and ideas of open science or maybe the open science movement is completely new to you. Whatever your current understanding is, we believe that what you learn here will be interesting, thought-provoking and useful in your future research career.

We are a group of budding researchers and PhD students who first met during a course focusing on open and collaborative research designed under the project Opening Doors funded by the Horizon 2020 EU programme for research and innovation. We felt that what we learned was both fascinating and helpful and believe that other students should have an opportunity to get familiar with these concepts too. Hence, we decided to create Agape. Agape means wide open, as is the open science philosophy and practice we want to promote. The word *agapē* originates from Greek and means love that is unconditional, such as our love for science (Okay, maybe not entirely unconditional, but you get the drift!). Under Agape we aim to share open science with fellow students and researchers, starting with this course and continuing with a series of workshops where we can learn, exchange our opinions and experiences, and together contribute to a better, more open future.

With this course, Agape would like to open doors for you into the world of open science and introduce various concepts that we think are crucial to high quality research and prior to the opening doors project we were entirely unaware of. Sure, we all heard about scientific integrity and open access publishing at some point in our studies, but the domain of open science encompasses a much larger set of ideas and concepts. Given the sheer extent of open science, this course does not, and could not cover the whole scope of open science. However, we will provide you with a scaffolding to understand the core concepts and signpost you to useful links and resources should you wish to delve deeper and start practising open science in your own work.

And now, without any further delay, let's quench that thirst for knowledge!

Course structure and certificate of completion

The course is structured into chapters that are written to expand on various topics. We think that the order they follow is logical and the latter chapters build on knowledge acquired in previous chapters. That said, you can decide to go through them in whatever order you like by clicking on different chapters in the menu on the left or to return to some of them should you find something is not clear or you require a brief refresher.

At the end of each chapter you will find activities to enhance your understanding of the concepts introduced in a particular chapter and to improve your practical knowledge. All chapters except one contain a short quiz consisting of five questions where you can test your freshly acquired knowledge. You will see your score immediately and you can save your high scores if you wish. These scores will be saved in cookies in your browser and you can delete them by deleting cookies. The quizzes have no time limit and you have as many attempts to practice as you want.

Once you read all chapters and practise each short quiz you can attempt the final quiz. If you score 90% or higher a certificate of completion will be generated for you and you can download it directly. You have unlimited attempts to successfully complete the final quiz but only three minutes for each attempt.

We know we are by no means perfect. We would love you to share your opinions, concerns, or feedback about a specific chapter or the course as a whole. There are two ways to do this. You can either fill in the survey or use the Disqus widget that you can find at the end of each chapter to access the forum. In a truly open spirit we will discuss, collaborate and offer constructive criticism and helpful advice and ask for the same from you. We will do our best to address your comments or pass them on to the course admins. We value all suggestions and feedback, and together we will make this course awesome. We just ask for a little patience.

Course evaluation

We have surveys at the beginning and end of the course to offer the best course experience possible.

Pre-course survey

You are invited to participate in a survey that will help the Agape team learn more about learners. We'd like to ask you to answer a few questions about yourself. Your honest feedback is very important for us to improve and serve

all learners better. Whether you simply browsed or completed the course, your feedback is valuable.

Thank you for taking this survey. Please click the link below to take the survey.

Pre-course survey link

Bug Report/Feature request

If you come across any glitches or bugs, or would like to see a different topic or feature in our MOOC, please open an issue in github or Bug Report/Feature request form

Contacts

Should you experience any technical problems or should you wish to share your ideas on how to improve this course email us on agape.open.science@gmail.com.

To share your thoughts and experiences either with this course or on open science in general, or to see what's new we will be delighted if you start following us on

Facebook Agape Open-Science,

Twitter @AgapeOpenSci,

Instagram Agape.Open.Science,

or on LinkedIn Agape Open Science.

Credits

Aswathi Surendran : content creator, IT whisperer
https://orcid.org/0000-0002-8709-6417

Cassandra Murphy : content creator, social media wizard
https://orcid.org/0000-0003-1332-359X

Ciarán Purcell : proofreading and editing rockstar
https://orcid.org/0000-0002-4376-599X

Marco Prevedello : content creator, IT advisor
https://orcid.org/0000-0002-8329-6294

Mohammed Mahmoud : content creator
https://orcid.org/0000-0002-1224-0381

Nina Trubanová : content creator, deadline overlord, vision pusher
https://orcid.org/0000-0001-8156-3304

Philipp Junk : content creator
https://orcid.org/0000-0002-5228-3896

Rasaq Semiu Abolore : content creator
https://orcid.org/0000-0001-6486-4754

Tendai Mukande : content creator
https://orcid.org/0000-0002-0654-7141

Una Ruddock : proofreading and editing balladeer
https://orcid.org/0000-0001-9118-4121

Wei Qi Koh : Info graphics wizard
https://orcid.org/0000-0001-8196-1628

Yao Zhang : content creator, social media wizard
https://orcid.org/0000-0003-0093-3882

And a big thanks to our muse Dr Denise McGrath.

How to cite this material

Agape Initiative. (2023). Agape Open Science MOOC. OSF. https://doi.org/10.17605/OSF.IO/DTB7V

Disclaimer

Any views or opinions represented in this course belong solely to the Agape team and do not represent those people, institutions, or organizations that the authors may or may not be associated with in a professional or personal capacity unless explicitly stated.

The information in this course is provided without warranty. The authors and Agape team have neither liability nor responsibility to any person or entity related to any loss or damages arising from the information contained in this course.

Chapter 1

Open science

Let's start at the beginning.

Open science is a movement to make scientific research, data and their dissemination available to any member of an inquiring society, from professionals to citizens.

Open science comprises several themes from conception to dissemination of knowledge. Based on principles of scientific growth and public access, open science includes practices such as open publishing and campaigning for open access, with the ultimate aim of making it easier to publish and share scientific knowledge.

1.1 What is open science?

Open science refers to a vision to improve scientific practices for reproducibility, transparency, sharing and collaboration of knowledge. Multiple pathways to achieving this vision have developed since the concept emerged in 1985 (Chubin, 1985).

As part of the global open science community, we expand the term "science" beyond its common use, such as life sciences or engineering, to include the arts, humanities and any other scholarly activities. Open science, open research and open scholarship are often used interchangeably. For consistency and ease of understanding, we use open science in this course.

1.1.1 Open science = open research = open scholarship

Generally, the open science movement identifies increased openness of scientific content, tools and processes as the key means of action. However, a single defi-

nition cannot encompass the diversity of the open science movement. Therefore, we give some alternative definitions of the phrase **open science** throughout this chapter.

The first definition we review is from the European FOSTER project. According to the FOSTER team, "open science is the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods." This definition highlights that open science is the act of improving access and contribution to all aspects of scientific practice: from research design, methodologies and tools, generated data, reporting and evaluation.

In an attempt to capture the complexity of open science, Fecher & Friesike (2014) define it as "an umbrella term encompassing a multitude of assumptions about the future of knowledge creation and dissemination". The authors summarise the movement complexity by identifying five schools of thought. Namely, the democratic school (concerned with knowledge access), the public school (concerned with accessibility to knowledge creation), the measurement school (concerned with alternative impact measurement), the infrastructure school (concerned with the technological architecture of science) and the pragmatic school (concerned with collaborative research). This (somewhat arbitrary) separation highlights the various paths in which science can be "opened".

Based on a review of published definitions and information Vicente-Saez and Martinez-Fuentes (2018) concluded that "open science is transparent and accessible knowledge that is shared and developed through collaborative networks". Here, knowledge includes code, data, ideas, information, scientific outputs, scientific publications and scientific results. Paic (2021) identified emerging trends in open science such as alternative reputation systems, open notebooks, open lab books, science blogs, collaborative bibliographies, citizen science and open peer-review.

Another definition we would like to present is the United Nations Educational, Scientific and Cultural Organisation (UNESCO) definition. Drafted at the 40th UNESCO General Conference in 2019 and officially published in 2021, the Recommendation on Open Science contains the following statement: "For the purpose of this Recommendation, open science is defined as an inclusive construct that combines various movements and practices, aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone. It also aims to increase scientific collaboration and sharing of information for the benefit of science and society and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community. It comprises all scientific disciplines and aspects of scholarly practices, including basic and applied sciences, natural and social sciences and the humanities and it builds on the following key pillars: open scientific knowledge, open science infrastructures, science communication, open engagement of societal actors and open dialogue with other knowledge systems."

Here, the UNESCO members clarify the meaning of the term science to include the humanities and the liberal arts. Furthermore, this definition highlights how the open science movement broadens its areas of influence beyond general research practices to the researcher community and society at large. The movement thus aims to better the inclusion of diverse ethnicities, cultures, languages, backgrounds and availability of resources across the scientific community.

Lastly, The Turing Way Community (2021) illustrates open research and its subcomponents as fitting under the umbrella of the broader concept of open scholarship in their handbook on reproducible, ethical and collaborative data science. These subcomponents are open data, open-source software, open-source hardware, open access, open notebooks, open educational resources, citizen science, equity, diversity and inclusion.

Many other definitions of open science can be found on-line or within written records. However, in consolidating the mentioned sources together, we conclude that six unifying principles and core values characterise the open science movement. These are:

- Transparency, scrutiny, critique and reproducibility

- Equality of opportunities

- Responsibility, respect and accountability

- Collaboration, participation and inclusion

- Flexibility

- Sustainability

1.2 The history of open science

Previously, we mentioned the vision and principles the open science movement shares. To better understand the implications opening science might have, it is useful to appreciate how the contemporary values and principles of both science and open science came to be. As Watson (2015) suggests, science is easily perceived as already "open", as already belonging to everyone. However, contemporary science is much more recent than one may think and its organisation has greatly transformed over the last 500 years. Even the current dominant practices of science are not immutable and they are likely to continue to change again in the future, evolving within the broader context of society.

1.2.1   17th to 20th century: The emergence of contemporary science

> "For in the sciences the authority of thousands of opinions is not worth as much as one tiny spark of reason in an individual man."
>
> – Galileo Galilei, ca. 1597

Between the 17th and 20th century, science underwent multiple reformations. If you were to meet scientists from before the 17th century, their perception of science would probably shock you. For centuries, they had more or less accepted the authority of the Church and of monarchs and their claims and theories didn't need to be backed up with either observable proof or the proof of reason. During the reformation, profound advancements in science took place and the scientific community gradually adopted scientific publications, formal review processes, scholarly associations, public grants and many more of the common modern practices.

The advent of printing and publishing companies in the early 17th century made public libraries more and more common. Knowledge started to accumulate in printed reports and encyclopaedias and, for the first time, it became accessible to the general public. At the time, this meant mostly privileged male and white citizens with greater socioeconomic resources and greater perceived standing in the social hierarchy.

During the same period, scholars started to move from unstable aristocratic patronages to assembled academies of science, slowly adopting academic publishing. Merton (1963) tells us that between the 1650s and the 1850s, the number of simultaneous discoveries ending in disputes dropped from 92% to 33%. Cryptic monographs and academic duels slowly became a thing of the past and science gradually became more open.

> "The assumption that peer review is as old as journal publishing [...] is based on a misunderstanding of Philosophical Transactions' editorial practice. [...] Indeed, for most of the history of scientific journals, it has been editors – not referees – who have been the key decision-makers and gatekeepers."
>
> – Aileen Fyfe, 2015

After the establishment of scientific publishing and academic societies, the development of a formal review process developed between the 19th century and first half of the 20th century. We might take today's practice of **peer-review**, in which one or more people with similar competencies (peers) as the author review a manuscript before publication on a voluntary basis, for granted. However, it took until the 1970s before peer-review became widespread. (Fyfe, 2015).

The first recognized formal review practice was implemented by the British Royal Society in 1832. A special committee within the society was responsible for accepting or rejecting submissions for publication based on independently written evaluations. George Gabriel Stokes, secretary of the Royal Society from 1854–1885, further refined this practice by sharing the referees' suggestions with the authors and facilitating the discussion between authors and referees. Similar review processes started to become common practice in other academic societies across the world during the 19th century.

On the other hand, private publishers would only introduce formal review practices in the 20th century and the journal editor(s) were the sole judge of acceptance or rejection of a submission. These trivial selection practices allowed for a fast research-to-publication cycle, making private journals appealing for swift communication between scientists. At this stage, societies' journals gradually hosted fewer and more refined publications, while private journals were the preferred communication channel between scientists.

1.2.2 Modern science and the fight for openness

As submissions grew in volume, private journals started to implement review practices in the second half of the 20th century, gradually bringing us to the current state of research practices. However, with private publishers flourishing, new economic and structural barriers to scientific knowledge arose. For example, nowadays a yearly subscription to a scientific journal can cost between 3,000 and 7,000 USD (Bosch *et al.*, 2022) compared with much more affordable mainstream media subscriptions. For example, the New York Times offers a yearly subscription for 20 USD.

Much of today's scientific publication is in the hands of five large publishing companies: American Chemical Society (ACS), Elsevier, Springer-Nature, Taylor & Francis and Wiley-Blackwell, who own 50 to 70% of the scientific writing market (Puehringer et al., 2021). A sentiment of anger is often apparent in scientific communities towards the private publishing sector. For example, Buranyi (2017) explains that "Scientists create work under their own direction – funded largely by governments – and give it to publishers for free. The publisher pays scientific editors who judge whether the work is worth publishing and check its grammar, but the bulk of the editorial burden is done by working scientists on a volunteer basis. The publishers then sell the product back to government-funded institutional and university libraries, to be read by scientists – who, in a collective sense, created the product in the first place."

The roots of open science stem from philosophical discussions ongoing in the 1970s and 1980s about what it means to have freely available scientific knowledge and not in the fight for more just publication practices (Chubin, 1985). However, we believe that the struggle for open access publication spurred the rise of the open science movement, which then drew together all of the elements previously mentioned.

A second steppingstone in the history of the open science movement was the advent of the Internet. From its infancy, the Internet shifted how information is shared and facilitated new methods of scientific exchanges. Literature research moved from public libraries to online repositories such as the Directory of Open Access Journals (DOAJ) which was launched in 2003 with 300 open access journals. Scientists also began to share early versions of their work or pre-prints through servers such as Biorxiv. Furthermore, online collaboration facilitated easily accessible information resources such as Wikipedia which became a household staple. A new reformation of science had begun.

> "The question is no longer whether open science is happening, but how everyone can contribute to it and benefit from this transition.".

– Audrey Azoulay, Director-General of UNESCO, 2021
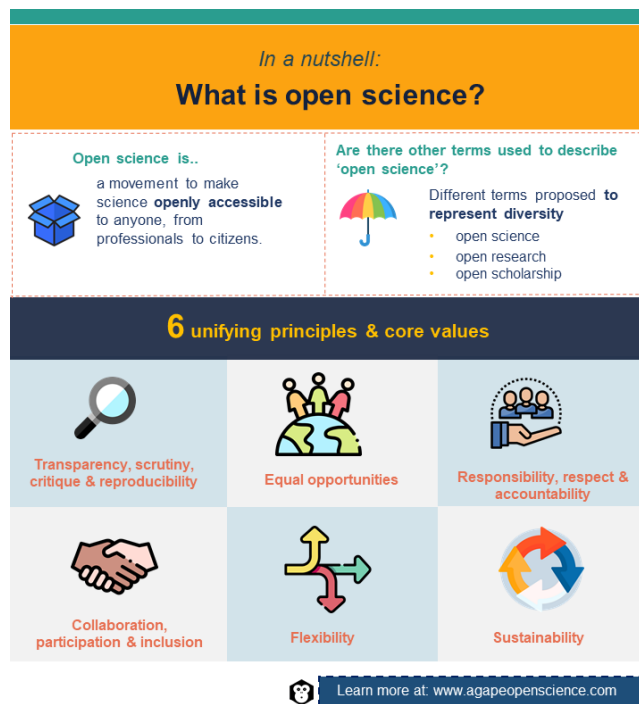
1.3   Test your understanding

Loading...

Activities

In a recommend activities section like this one, we will recommend the activities to increase your understanding of the concepts and improve your practical knowledge.

- When did you first hear about open science?

- Propose one change you could bring to your workflow to implement one of the core values of open science presented.

- Try to imagine how open science will change during your lifetime. And what about the 22nd century? What do you think the future of science will look like? Will it finally become fully open?

- Share anything interesting that you learned or found in this chapter with others on our social media.

Chapter 2

Open data and open access

Now, let's have a look at open data and open access. What exactly is it?

2.1 Definitions and history

"Open data refers to data access and sharing arrangements, where data can be accessed and shared and reused by anyone without technical or legal restrictions, free of charge (to the greatest extent possible) and used by anyone for any purpose subject, at most, to requirements that preserve integrity, provenance, attribution, and openness" (OECD, 2015). Access to data can occur along a spectrum, with "different degrees of openness, depending on the community of stakeholders involved. 'As open as possible, as closed as necessary' is often used to illustrate the fact that while opening up data can help advance the science, technology and innovation (STI) agenda, this needs to be balanced against issues of costs, privacy, security, intellectual property rights and preventing malevolent uses" (Paic, 2021).

According to the Open science monitor of the European Commission, the open science scholarly community, open access to publications and open research data comprise the main pillars of open research and open science.

Without open data, scientific research would progress very slowly. The idea of open data stemmed when large international consortia collaborated on complex projects and sharing data became a necessity. You may already be thinking about some of these applications such as oceanography, particle physics, molecular biology or genetics projects. The first significant initiative in data sharing is considered to be the World Data Center, established in 1957 by the International Council for Science. Its original purpose was to serve the International Geophysical Year, a worldwide effort to study the Earth, oceans and atmosphere

in a coordinated and synchronous way (Korsmo, 2010). Data from many different fields can and should be made open. These include but are not limited to science, finance, business, government, culture, weather and the environment.

It makes sense that research data and scientific publications funded by taxpayers should be open and accessible to the public for free. Arguably, the positive societal impact of free access to scientific knowledge outweighs the investment of time and money. The European Commission identifies open science and open access to data as the dominant driver for the future of science, with high expectations of improved scientific integrity, a better connection between science and society and making science more responsive to societal challenges. You can learn more about the benefits of open data and exceptions to this rule in the chapter "Research data". Whilst some of the ideas around open science and open access are really amazing, it also comes with some negatives. You can learn more about this topic in the chapter "Pros and cons".

But let's get back to the topic of this chapter. Where does open access come from and how can it be defined?

Machado (2015) traced the origins of open access to research data that started as anonymous transfers through file transfer protocols (FTPs) within some private networks and exchanging of physical media, such as tapes and disks. This gave rise to the first free access databases of electronic open access data, the Educational Resources Information Center (ERIC) and Medlin (NLM) which are managed by the National Library of Medicine and the National Institute of Health now known under the name PubMed, in 1966 in the USA. These were followed by other catalogues of scientific literature and books. Then everything changed in 1990 with the arrival of the internet. Its concept was developed by the European Laboratory for Particle Physics (CERN) as an answer to the ever-increasing needs of particle physicists to exchange large volumes of data (Berners-Lee *et al.*, 1992). In the following year, the repository of physics, mathematics and computer science texts arXiv was created, followed by the genetic research database Genbank in 1992. Since then, databases and repositories have played a key role in open access, allowing the availability of articles, papers and research materials produced by universities and research centres.

What exactly do we understand regarding the term open access? According to the OECD (2015), open access is "unrestricted online access to scientific articles, via a number of channels, such as institutional repositories, journal publishers' websites, researchers' webpages, etc."

2.2   Types of open access

Open access (OA) publication means making a publication freely accessible online in a digital format with no barriers to access. There are different types or levels of open access:

- **Green open access** – Articles where the author or institution provides access. This is often referred to as self-archiving. Usually, researchers submit their manuscript (published or sometimes unpublished) to an archive. Most institutions provide open access archives. This is usually where preprints are published while undergoing review. Preprints are full drafts of research papers which are currently under peer-review for publication. Researchers choose to share these publicly prior to a completed review to allow for feedback and visibility of their results while they wait to hear back from the journal. It is essential to ensure that the authors comply with the publisher's copyright policy. Some publishers have a standard embargo period before making the work openly accessible. Even in that scenario, meta-data is often exempted from this restriction.

- **Diamond or platinum open access** – Journals do not charge either readers or authors directly. Publishers then often require funding from the government or non-for-profit, non-commercial organisations, associations or networks or rely on advertising. The peer-review process is performed by volunteers.

- **Gold open access** – Immediate access to an article and data upon publication is provided by a publisher. Publishing costs can be recovered through fees, but more often, an article processing charge is covered by the author, institution or the funding body of which the research is being sponsored. Gold open access provides a rigorous peer review mechanism. More information on publishing gold access can be found on the PhD on Track website.

- **Bronze open access** – The article is free to read only on the publisher's page. However, it lacks the open licence for reuse.

- **Hybrid open access** – This is a controversial model in which an author or institution is required to pay the open access article-processing charge to make their paper available open access in a traditional journal which provides a subscription service. It is advised to stay away from this type of open-access publishing as many institutions and funders will not agree to pay a fee.

- **Black open access** – Free access to publications behind the paywall when people with access share free copies. This is an unauthorised large-scale copyright infringement. Black open access can take the form of shadow libraries, such as Sci-Hub or Library Genesis. This may also be done by sharing the publication via social media, such as with #ICanHazPDF hash-tag on Twitter.

Budapest Open Access Initiative defines the terms "gratis" and "libre" in order to distinguish between free to read versus free to reuse. Gratis open Access (⊘) refers to an online access to read the article free of charge. Similarly, libre open

access ( ) refers to an online access to read an article free of charge, however this includes some additional rights to reuse the article under specific Creative Commons licences. Libre open access covers types of open access defined in the Budapest Open Access Initiative, the Bethesda Statement on Open Access Publishing and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities.

2.3   Quid pro quo

What are the benefits of publishing your studies in open access journals?

Citation is still considered an important metric of influence and impact of academic work and open-access publications are proven to have a higher citation impact (Langham-Putrow *et al.*, 2021). Open access can also help in other ways. For example, it makes publications and data available to researchers from countries and organisations who cannot afford to pay for access. This allows greater access to current scientific research which can help connect people to find solutions to problems they are targeting. For instance, tackling increasing crop yields to address growing populations during the battle against climate change requires lower socio-economic nations and regions to contribute and collaborate and open access facilitates this.

In order to make data open, it needs to be compliant with legislation, such as General Data Protection Regulation (GDPR) in the EU or cyber security legislation, and follow FAIR Guiding Principles for scientific data management and stewardship first published by Wilkinson *et al.* (2016). Processing personal data transparently makes publishing and re-use of open data easier and patients or respondents more involved in the research. You will learn more about these principles in the chapter "FAIR principles" and about policies in the chapter "Open science policy, scientific integrity and ethics".

The number of funding agencies and research institutions adopting open science policies is increasing every day, with Europe being in the lead, followed by North America, Asia, Latin America, Oceania and Africa (ROARMAP, n.d.). Nowadays, about half of published papers are open access and the number keeps increasing. However, it varies greatly based on scientific discipline. In a study of articles published between 2009 and 2015, more than 80% of astronomy, astrophysics, embryology, tropical medicine and fertility papers were available in open access. In contrast, less than 10% of those in pharmacy, applied, inorganic and nuclear chemistry and criminology were open access. This study also unearthed that the dominant category of open access is not green or gold open access, but articles made free to read on the publisher's website, without an explicit open licence (Piwowar *et al.*, 2018).

We live in an era where data-driven innovation is transforming society, and the need for open science is unanimously recognised in the scientific research com-

munity. Governments, funding agencies and institutions recognise the benefits of open science and open research and are steadily developing strategies, policies and guidelines as groundwork for its existence. Since the original OECD Recommendation was adopted in 2006, many of these policies have been implemented at national and institutional levels and have contributed to significant advancement in this area and open science and open data have become mainstream. At least 58 countries have adopted dedicated national strategies and policies for open data and publications (EC/OECD, 2018). This has had a significant impact in areas such as the reproducibility of scientific results, diffusion of knowledge across society, cross-disciplinary co-operation, resource efficiency, productivity and scientific advancement. An updated OECD Recommendation on research data published in January 2021 (OECD, 2021) emphasises the relevance and importance of several key principles set out in 2006. These are openness, flexibility, transparency, legal conformity, protection of intellectual property, formal responsibility, professionalism, interoperability, quality, security, efficiency, accountability, and sustainability. It also expands the scope to cover not only research data but also related metadata, bespoke algorithms, workflows, models and software that are essential for their interpretation.
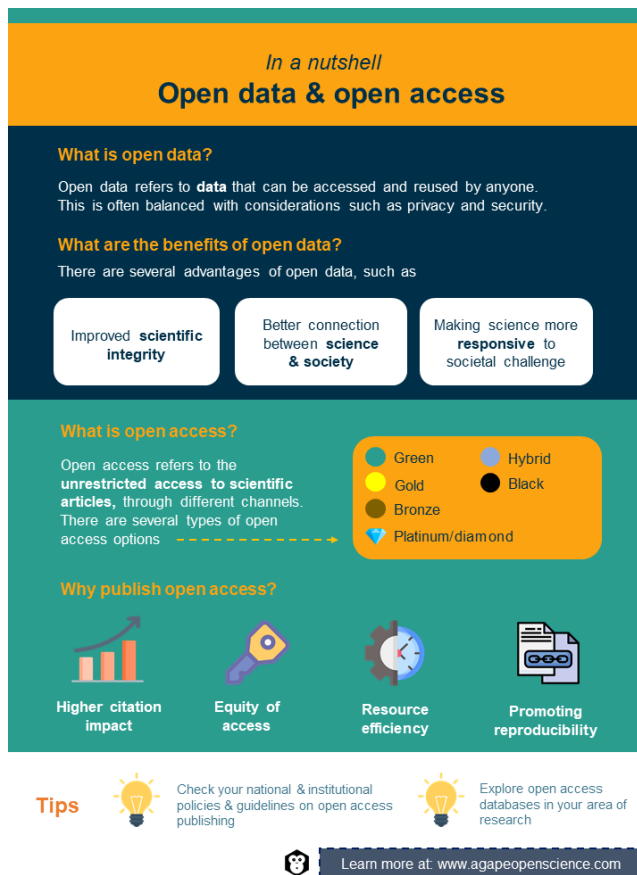
2.4   Test your understanding

Loading...

Activities

In a recommend activities section like this one, we will recommend the activities to increase your understanding of the concepts and improve your practical knowledge.

- If you're looking for a tool that enables you simple, free and legal open access to research articles try Open Access Button. Check it out!

- Explore a few databases, ideally those that are close to your area of research. You can also check some of those mentioned in this chapter. Look at some data and explore how these databases work.

- If you are an Ireland based researcher, you can check out the list of Open Access Journals that are close to your area in the Irish Open Access Publishers (IOAP) website.

- Check your national and university policies and guidelines about open access publishing. It is possible that your university also has an agreement in place that allows its corresponding authors to publish an agreed number of open access articles without paying the processing charge. Can you find their list and identify those journals where you could publish your research? Discuss this possibility with your supervisor/PI.

- Share anything interesting that you learned or found in this chapter with others on our social media.

Chapter 3

Open source, open licencing, scientific programming

3.1 Open-source

It is no secret that the open science movement got its inspiration from the open-source culture movement. Open-source software refers to source code that anyone can inspect, modify and enhance because the licence under which it is released grants permission to do so. You can find a more detailed definition on the website of the Open Source Initiative. Similarly, Open Source Hardware Association defines open-source hardware as "hardware whose design is made publicly available so that anyone can study, modify, distribute, make, and sell the design or hardware based on that design."

The main benefits and reasons for the adoption of open-source software and hardware (Casson and Ryan, 2006) can be categorised as follows:

- Security
- Affordability
- Transparency
- Perpetuity
- Interoperability
- Flexibility
- Localization

3.2   Open licensing

Even when your data, software or hardware design are made freely available in the public domain, an explicit licence would provide legal clarity on the access and re-use of it. You can licence the data only if you are the rightful owner. Licensing helps to:

- Remove the ambiguity on the re-use of data
- Exempt users from copyright infringement
- Ensure that the source author is credited rightfully
- Ensure that the re-used or re-distributed data remain open access
- Ensure that data is not misused or distorted

How to choose a licence

1. Ensure that the data is copyrightable. This may vary across domains, jurisdictions, funders, etc.
2. Check the licensing obligation of the funder(s), institution(s), government, data centre or repository.
3. If your work is a derivative of a third-party author, ensure to comply with the source data's licensing requirements.
4. Select the data licence with the conditions that meet your criteria and that covers the content that you want to share.

The most common conditions found in data licences are:

- **Attribution** (BY): The source/author must be acknowledged when it is distributed, displayed, performed or used to derive a new work. If you are using data from multiple sources, each contributor needs to be acknowledged.
- **Copyleft** or **share alike** (SA): Any new work derived from the licensed data should be released under the same licence of the source data.
- **Non-commercial**: This type of licence prevents the user from using the data for commercial purposes.

3.3   Licence providers

Prepared licence: Research Institutions or other data publishers can create licences. For example, the UK Data Archive requires that you sign a standard licence agreement that clarifies the rights and responsibilities of both parties and permits the UK Data Archive to perform its curatorial function.

Bespoke licence: If the existing licences don't meet the author's requirement or cater to special circumstances, they can make their own licence. In this scenario, it is mandatory to ensure that the custom licence complies with any existing legal bindings.

Creative Commons: One of the most popular and widely accepted licence providers for most content with the exception of source code. Three versions (CCO, CC-BY, CC-BY-SA) of it are intended for open licensing. Choose a licence by GitHub provides a list of licences that are specific to software codes.

Open data commons: Similar to Creative Commons, but these licences are specifically designed for databases.

3.4 Scientific programming

Documenting your code

To make your research open, i.e. transparent and reproducible, it is good practice to share not only your data but also the code used for the analysis, modelling, visualisation, etc. This code is then known as open-source research software. When sharing your code or software, it's good practice to also include documentation explaining how to use your code. So why do you need to prepare this documentation?

Benefits for you:

- In six months' time or whenever you choose to work on it, you'll still be able to use your code.
- You want people to credit you when they use your code.
- You want to learn how to be self-reliant
- You may attract others to contribute to your code. Benefits for others:
- Others can simply utilise and extend your code. Benefits for science:
- You are contributing to science.
- You are promoting open science.
- Documentation allows for clarity and reproducibility.

Here are **best practices** for writing documentation:

1. Provide a README file with the following information:

   A quick overview of the project

   Instructions for installation

   A brief example/tutorial

2. Allow others to use the problem tracker.

3. Create application programming interface (API) documentation.

4. Write down your code.

5. Use coding conventions, including file structure, comments, naming conventions, programming methods, etc.

6. Include an introduction for contributors.

7. Provide citation information.

8. Include any licensing information.

9. Include a link to your email address.

10. List all the file versions and the fundamental changes you made.

A helpful hint: When naming files, make sure their names are descriptive and consistent!

Importance of scientific programming

There are multiple ways in which scientists and researchers can benefit from scientific programming. Scientific programming significance includes a wide range of abilities without focusing on any particular field. Generally, scientific programming can facilitate the following:

- **Time-consuming tasks can be automated** – Automating tasks using scientific programming can simplify long-term tasks or those that are impossible to do by hand. Imagine, for example, that you want to figure out how many tweets were posted about a recent natural disaster and you have to sift through tens of thousands of feeds one by one. A few minutes might be enough to complete this task with code.
- **Creating adaptable research** – You can modify and rerun your code repeatedly if you write it correctly. Consider you are researching the relationship between socio-economic data and air pollution in a particular location. Using a properly structured and well-commented script, updating each year's socio-economic data can be easily incorporated.
- **Help to publicise the research and share the findings with other researchers** – Because code is so easily accessible, research becomes more open and repeatable. It helps the researcher to convey their specific methodology to other experts as well as the community.
- **Documenting your thinking** – You can quickly document your strategy with code. You may use comments to describe each stage of the process (to your future self or others), making it quick and easy to update or adjust things afterwards.

- **Research collaboration** – Collaboration is facilitated by the use of code. Returning to the previous example, if you are researching air pollution in a particular location and a colleague is researching air pollution in another location, you may compare models, swap scripts and collaborate.

The above five features of scientific programming assist the researcher significantly in various ways and are considered key tools to nudge research forward. The significance can be simply identified by the speed of conducting the research through a high level of computation and, most importantly, the collaboration and the modifications that may apply. Scientific programming has been, and continues to be, ground-breaking. From assisting biologists in sequencing the human genome to allowing social scientists to make better economic forecasts the applications are limitless.

What exactly is scientific programming?

There is a simple definition of scientific programming, yet it covers a vast array of applications and industries. Using a computer-aided program for scientific research is referred to as science programming. Scientific programming can be useful for most scientists and researchers, especially PhD researchers. The rate and reproducibility of a researcher's work can be exponentially increased using scientific programming. Computers, designed for efficiency and scale, can perform massive calculations, store data and analyse results. By automating processes, scientists are able to save time and effort and make research more accurate, reliable and efficient.

It is essential to note that computers are error-free when it comes to mathematical processes. Occasionally, mistakes can happen, but these mistakes usually occur because people make errors when using computers. Computers follow directions, so if a calculation goes wrong, the computer will not understand it independently. However, a computer can do calculations within minutes that would take researchers months or even years to perform. Furthermore, the code will execute the calculations consistently for each run.

Respond to disasters: an example of the power of scientific programming

In addition to its many advantages, scientific programming can accomplish a great deal of work that would be impossible for one person or even a team of people to complete without it. Lise St. Denis's research on climate change at Earth Lab demonstrates that power clearly. Lise uses Twitter to notify first responders about emergency situations that result from natural disasters as they develop or progress. Without this technology vital time sensitive information may be missed.

In the event of a disaster, the police are contacted alongside emergency services. Disaster survivors also reach out to their online communities, sometimes providing vital information. The call volume of hotlines during disasters can be overwhelming for authorities and the only way for people to communicate may be through social media. Thus, sites like Twitter can be a great source of information for disaster response teams. However, the volume of tweets makes it difficult for one person (or team) to vet all the information and still get it to emergency response teams in time. Lise St. Denis witnessed this through her extensive experience of natural disasters. For instance, during the Carlton complex fire of 2014, she worked on a team that was tasked to sort tweets and compile a full report. Although Lise's team provided useful information, it was unable to keep up with the volume of tweets and responders needed the information faster than the team was able to provide it. For Lise, the answer was obvious, they needed to deploy the superpowers of scientific computing.

Since the beginning of 2013, Lise had been developing a filtration algorithm to harvest data from Twitter and sort it by importance. Using this code, one person can automate the work of a whole team of humans by analysing and categorising every single tweet as they arrive. As a result of the algorithm, tweets are separated into those which first responders need to know about and those which the algorithm deems as less important. One of scientific programmings' capabilities is the ability to "look" at enormous amounts of rapidly generated data and categorise it.

The future of scientific programming

Because there are so many fascinating possibilities, it is difficult to pick one field of scientific programming that is remarkably promising. Almost every discipline of study has a programming tool that could be considered as the "future of programming" and it would be difficult to discuss and list them all.

Modelling is a promising development that serves multiple professions. For decades, models have served as the foundation of science. There are a wide range of examples of using modelling techniques in science, ranging from Earth science (predicting wildfires) to medicine (analysing illnesses). No model is perfect. Thus, there is continuous development in the pursuit of better, more precise models with complete algorithms producing reliable results.

In today's data-driven world, the term science is closely linked with scientific programming. Problems that have baffled scientists for decades are addressed in a matter of seconds by leverage the power of large computers. The rise in efficiency and speed has completely transformed most sectors of modern science. Without question, scientific programming is our way forward. To be part of this movement, you can share your code with others and always appropriately cite the open source you use.

3.5 Test your understanding

Loading...

Activities

In a recommend activities section like this one, we will recommend the activities to increase your understanding of the concepts and improve your practical knowledge.

- Learn more about open-source software and hardware on Open Science Training Handbook or try to work on a DIY open-source hardware project.

- Have you ever checked licences on the types of software you are using? Now is the time to do that. Is the software you are using open-source software?

- Working on your research project can be exhausting at times. Why don't you try relaxing and practice open-source software skills by playing some of the following games? Play and learn!

  CodeCombat: These games take you step by step through ideas, starting with basic computer science and gradually increasing in difficulty.

  CodinGame: When you have a better grasp, this game is about solving challenges in specific languages.

  CodeWars: Get right into programming challenges and experience debugging your code.

- Have you ever shared your code, worked on an open-source hardware project or do you have any other experience related to this chapter? Share it with others on our social media.

Chapter 4

Pros and cons

4.1 Misconceptions about open science

> "Open science describes the practice of carrying out scientific research in a completely transparent manner, and making the results of that research available to everyone. Isn't that just 'science'?"
>
> – Watson, 2015

While the situation is improving, there are still many misconceptions about open science that give it a negative image. In 2014, a survey by the Nature Publishing Group and Palgrave Macmillan found that 40% of scientists that had not published open access said "I am concerned about the perceptions of the quality of open access publications". Other misconceptions commonly cited include: "There is no peer review for open access publications", "Open science leads to worse research" and "In open science, it is possible for others to steal your research". We believe that such misconceptions arise from a lack of understanding of open science. Two motivations in creating this course are to demystify open science and challenge these misconceptions.

Encouragingly, there is a positive trend in understanding open science and acceptance within the scientific community. In a 2015 follow-up survey, only 27% of scientists expressed their concern with the perception of the quality of open-access publications compared to 40% in 2014.

4.2 Cons

While we have created this course because we believe in open science, there are currently some issues with open science that we feel obliged to discuss:

- **Concerns about quality** – While we will talk about the improvements that open science brings to the world of science later, there are some legitimate concerns. Preprints accelerate open science but make science accessible before the peer review process. The reader has to be more careful when reading and interpreting preprints. However, most manuscripts on preprint servers are considered to be final drafts of manuscripts that are/will soon be submitted to journals for peer review and processing and should therefore meet high quality standards. Similarly, predatory journals undermine the quality standards expected from scientific publishing. You, the scientist, need to be aware of the phenomenon and pay attention to identify predatory journals when reading or publishing (see Beall's List).

- **Time- and effort-consuming** – Many open science practices are both time and effort-consuming. For example, after publishing open-source software, you might need to spend time on bug fixes and updates, or on interacting with potential users and updating your documentation, all at no further benefit to you. Properly annotating and publishing data sets takes a considerable amount of time. Sometimes, going the open science route can even be associated with additional financial costs, such as publishing under an open access licence. Sadly, there is no good way to sugar-coat this: participating in open science will take up some of your time. However, we will shortly discuss the incentives for spending your time (your most precious resource during your research journey).

- **Open science is not properly incentivised** – Currently, open science practices are not properly awarded and incentivised. The hard currency of high impact publications is still one of the most valuable aspects of your CV. Devoting time and effort to open science aspects that do not contribute to this metric can seem like a waste of time. However, the situation is getting better. More and more funding agencies are beginning to require open science practices in their projects and additional metrics apart from the number of publications and citations are becoming increasingly popular. That said, currently open science practices are not as highly valued as generating more traditional research outputs such as journal articles.

4.3   Pros

On the other hand, there are also plenty of positive aspects of open science.

- **Individual benefits for your research** – There are collective benefits of open science that can also be useful for your research. As we have already established, practising open science requires a considerable effort from the individual for the benefit of the collective community. However,

as part of the community, you can also access, and benefit from, the efforts of others. Maybe there is a piece of software that you can use because someone made it available open-source. Maybe there is an interesting dataset that is publicly available and relates to your research question. There are also many big collaborative resources that are produced under the umbrella of open science. Open science can produce opportunities for your research that "closed" science could not.

- **Investment in your career** – One of the trends in research we have already touched upon is that the general trend in the scientific community is towards adopting open science. A good example of this is open access publishing, which is increasingly required by research institutions and funding agencies. Similar trends can be seen with regards to data deposition or code sharing. Starting to incorporate these open science practices into your research as early as possible shows your ability to adhere to open science practices, which might give you an edge when applying for a position or funding. In addition, there are some more direct benefits to your research from open science. Open science, by definition, makes your research more accessible and visible. It has been shown that research articles published with an open access licence are on average cited more often (Langham-Putrow *et al.*, 2021). Other researchers can more easily engage with or try to reproduce your research or use and cite your data. While not all these achievements are captured by traditional metrics, they can be useful investments in your own career.

- **Open science aims to make science better** – One of the reasons there has been a move towards open science in recent years has been the replication crisis, in which it was found that the results of many scientific studies were not reproducible. One of the lessons learned from that was to move towards a more open, transparent and inclusive scientific environment. Open science was therefore designed to make science better. Full access to protocols, methods and analyses can make science more reproducible. Combating publication bias that exists in scientific literature by making data and negative results available can reduce the number of experiments necessary or even help to avoid pitfalls that others might have encountered before. It is well known that results of a treatment, drug or process that have a positive effect on the target group or issue tend to get published more than research that finds negative or no effects. This may be due to authors' motivation to publish but often it is due to journals and publishers being less willing to publish research that might not grab the reader's attention and therefore sell more copies. Pre-prints and open access publishing help to counteract this issue. Pre-prints can also help to accelerate science as well as improve manuscripts by open peer review. Here we can see the real-world impact of open science where both sides of the "scientific story" are shared and end users of treatments, drugs, etc. are better informed regarding the overall body of evidence and can make a more informed decision.

4.4   Conclusion

After discussing the positive and negative aspects of open science, we think it is fair to come to the conclusion that currently open science practices benefit science and the community in general, granted the benefits to individual researchers may not be immediately visible and an investment of time and resources is required. However, we strongly believe that the scientific community as a whole will move towards an increasingly open science approach in the coming years and decades. While currently there might not be enough individual incentives, these will come. Maybe in some disciplines and countries faster than in others.

Finally, hearing about all the best practices in open science can be intimidating. We think it is very important to point out that, while all these practices are important, we believe that it is best to figure out what is possible and appropriate for your work in particular and start implementing these practices bit by bit.
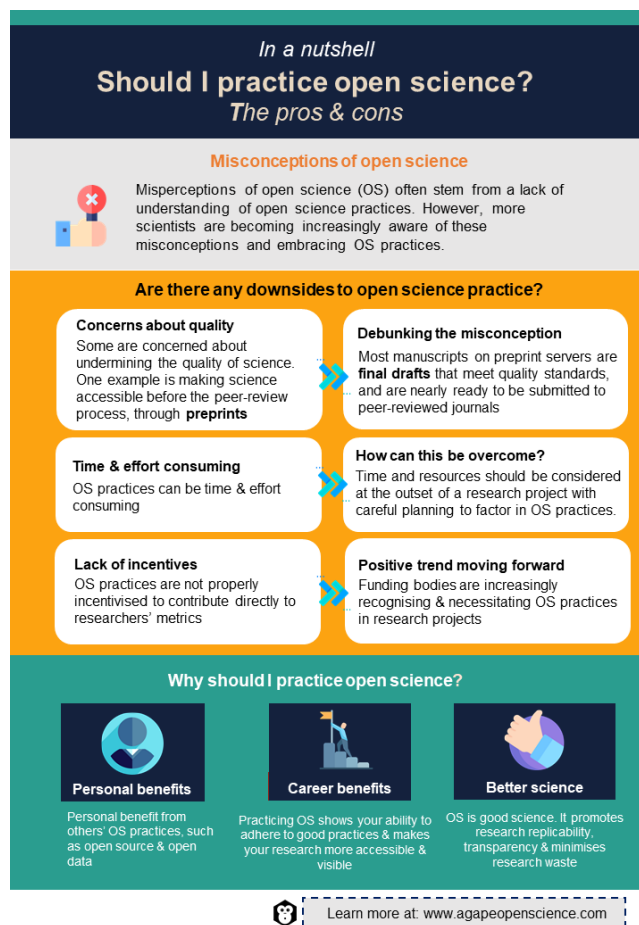
Activities

In a recommend activities section like this one, we will recommend the activities to increase your understanding of the concepts and improve your practical knowledge.

- Sort out your thoughts about open science by filling in the ORION Questionnaire.

- Think about how much time and effort you feel confident to dedicate to open science.

- Have you come up with other pros and cons that you think are relevant and were not mentioned in this chapter or the ORION Questionnaire? Is there anything you would like to share with others? Feel free to do so on our social media.

Chapter 5

Research data

Research data are a fundamental part of the research process and open science.

But what is research data?

The University College Dublin (UCD) Research Data Management Policy defines research data as "information collected to be examined and considered and to serve as a basis for reasoning, discussion or calculation. It is used as a primary source to support technical or scientific enquiry, research, scholarship or artistic activity, is used as evidence in the research process and/or is commonly accepted in the research community as necessary to validate research findings and results."

Research data exists in a specific research context and provides evidence or validation to claims and findings relevant to a specific research question. Data can be quantitative or qualitative and in physical, digital or analogue format. Quantitative data are measures of quantity and are recorded as numbers. In the research community, there are widely accepted standard units that allow specific quantities to be expressed in ways that are unambiguous and universally understandable. Qualitative data carries information about quality and do not take the form of numbers. Physical data refers to samples and specimens.

Examples of research data:

- Documents, spreadsheets and notes.
- Laboratory notebooks and field notebooks.
- Laboratory protocols, methodologies and workflows.
- Questionnaires, surveys, interviews, transcripts, codebooks and test responses.
- Standard operating procedures and protocols.
- Photographs, films, digital images, audiotapes and videotapes.
- Protein or genetic sequences.

- Spectral data.
- Slides, artefacts, specimens and samples.
- Maps and geo-spatial data.
- Collection of digital objects acquired and generated during the process of research and results of computer simulations.
- Database contents (video, audio, text and images).
- Models, algorithms and scripts.
- Contents of an application (input, output, log files for analysis software, simulation software and schemas).

Research data can be categorised as:

- **Observational** – data captured in real time, usually unique and irreplaceable. It is collected using methods such as human observation, open-ended surveys or through the use of an instrument or sensor to monitor and record information. E.g., weather data, noise level and recordings.

- **Experimental** – data collected through experiments or clinical trials by the researcher to measure change or to create differences when a variable is altered. It helps to determine a causal relationship and is typically projectable to a larger population. This type of data is often reproducible, but it can be expensive to do so. E.g., sequencing data and quantitative data recorded with laboratory equipment.

- **Simulation** – data generated using computer test models that try to determine what would happen under certain conditions. The test model and metadata can be more important than the output. E.g., climate predictions, economic models and chemical reactions.

- **Derived or compiled** – data originating from processing or combining existing data points, often from different data sources. It can be replaced if lost, but this can be very time-consuming and/or expensive. Typically used in secondary research. E.g., databases and population statistics.

- **Reference or canonical** – collection of smaller datasets, usually published and curated. E.g., IUCN Red List of Threatened Species and NASA Earth science data.

Next, we can divide data into primary and secondary. **Primary data** are collected or generated first-hand to answer a specific research question. **Secondary data** refers to existing data that is being reused for a purpose other than the one it was collected for. It tends to be readily available, include large samples and be collected over a long period of time. Nowadays, high quality research and publications can be made using only secondary data. Thanks to open science, sharing and access to data utilising this approach is becoming more popular. It can be, and often is, more cost-effective than primary research. One

potential drawback of secondary research is a lack of control over the research question, the data collected and the methods used. The potential strengths and benefits of secondary research do not undermine the importance of primary data collected in well-designed experiments and studies that are necessary for improving our knowledge and understanding of the world around us.

Metadata is another important type of data. It is often referred to as data that provides information, background or context information about other data. Put simply, metadata is data about data. In other words, metadata is structured reference data that helps to sort and identify attributes of the information it describes. Examples of metadata are author, type of data, file size, the date the document was created, HTML tags, geolocation, environmental conditions affecting the main variable and instruments used to collect or generate data. You can learn more about metadata in the chapter "FAIR principles".

Other types of data exist that are not commonly shared because of the nature of the records themselves or because of ethical and privacy concerns. E.g., preliminary analysis results, drafts of scientific papers, peer reviews, communication with colleagues or stakeholders. Research data also does not include trade secrets, commercial information, materials necessary to be held confidential by a researcher until data are published or similar information which is protected under law. Personal and medical information that could be used to identify a particular person or culturally sensitive data are special types of data that come under specific legislation. In the EU the relevant regulation is the General Data Protection Regulation (GDPR).

5.1 Research data life cycle

When working on your research, you are very much focused on your project carrying out the necessary practical tasks. The challenge is to also think about data management. It is in your best interest due to the long-term value of data you might be generating. You want to make sure that at the end of your research not only your data, but also metadata and documentation are complete, preserved and made accessible so that other people can use it and that you get credit for all the hard work you put into collecting or generating that data.

Data management is the practice of collecting, preserving and sharing research data. Data management continues beyond the duration of a specific research project and covers all aspects of curating and caring for data. Different activities and stages of this process that can be schematised in the research data lifecycle model include:

1. **Planning** – The first step is identifying data to be collected or generated in your research. It should include the nature, scope and scale of data. Resources and costs associated with data collection should be identified. This will depend on which methodologies or software will be used if

new data are collected or produced. Data preservation should be planned before data are collected or generated.

2. **Collecting or generating data** – For your research you can use either primary or secondary data. In the case of primary data, you are collecting, generating, storing and organising data and metadata. If you are reusing data made accessible by someone else, these are referred to as secondary data.

3. **Processing data** – Processing means converting raw data to formats suitable for analysis or generating new variables. Necessary steps are also cleaning and standardising data and applying quality controls. All steps of data processing activities, including scripts and outputs, should be documented.

4. **Analysing data** – Essential parts of your research are data analysis and interpretation. Statistical analysis, computational analysis and data visualisation are used to produce research outputs. All these steps must be reproducible. Therefore, it is important to document all steps of this process.

5. **Preserving data** – Data of long-term value should be preserved and made available for others to reuse. This involves selecting data for preservation, converting data to other formats, creating supporting documentation and depositing data in data centres, data repositories or institutional data repositories for preservation. It is important to plan data preservation from the very beginning of the research project in order to collect all necessary metadata.

6. **Making data accessible** – Creating online metadata records for data in a data centre/repository, obtaining a unique persistent identifier for data, licensing data for reuse, enabling access to data via a data centre/repository and citing and linking to data and code from research outputs all fall under making data accrssible.

7. **Re-using data** – For your research, you can use secondary data collected and made accessible by other researchers. Similarly, your primary data can be used as secondary data. It can be used by other researchers or by you to conduct secondary analysis or follow-up research, by policymakers to inform evidence-based policymaking or used by the scientific community in communication and engagement with the general public, industry, private sector and media.

5.2 Data management plan

Developing a data management plan (DMP) can be invaluable to your research by ensuring efficient research data management and sharing. A DMP is a docu-

ment that outlines how data are handled throughout the entire research process and once it is completed. It should consider all aspects of the research data lifecycle. A DMP should be produced before you start working on your research project and you should continually update and edit it when needed. A DMP is not a static but living document. It is okay if you can't answer all the questions at the beginning. Although it requires time to create your DMP and to keep it updated, in the long-term it saves a lot of time and effort. Also, public funding bodies often require at least an outline of a DMP. It is good to know that requirements for a DMP differ between various institutions and funding bodies. In general, it may provide information to answer the following questions:

- What is the nature of your research and your scientific hypothesis? What questions are you trying to answer?

- Who has what roles and responsibilities? Do you have any special requirements for hardware and software?

- How are you planning to generate or collect data? If working with physical samples, how will these be labelled and what system of unique identifiers will be used? If working with people or animals, how will ethical issues be handled?

- How will these data be processed?

- What quality assurance checks will be carried out and how will you deal with problems, missing values and errors, if found?

- How will data be stored and shared during the project (permission levels, version control and backups) and once it is completed (archiving)? How will intellectual property rights be handled?

Although not addressed in a DMP, other types of documents and records should be managed during and beyond the life of a project. E.g., correspondence (e-mail and paper-based), project files, grant and ethics applications and approvals, signed consent forms, research reports, technical reports, project reports and files and master lists.

5.3 Data access

Data can be stored in data centres or repositories. Sometimes, your university or institution might require you to store your data in its repositories. Make sure to familiarise yourself with policies and requirements when planning your research. You can learn more about Data centres and repositories in the chapter "Data repositories and data centres".

Not all data uploaded into data centres or repositories are necessarily accessible to everyone. Whether there are concerns that releasing some data can have negative consequences or you just need more time to publish results of analysis using those data, special restrictions can be applied on how others gain access to your data. Depending on the repository, access can be classified, for example, as:

- **Fully open access** – Open data has no restrictions on access. Anyone can view and download it. This makes it more likely to be reused, for others to verify the results of your research and for you to get credit for generating and publishing that data.

- **Embargoed access** – Embargoed data are made available within a specific period of time. Until then, only metadata is made public. At the end of the embargo period, data will become available by either open or mediated access, depending on the option that you've selected. This should give you enough time to publish your findings or to register a patent.

- **Mediated access** – Although metadata is made public, for someone to access your data you must approve an application that should meet conditions you have outlined. This may include requesting proof that a person asking for access to your data is a genuine researcher and that they have ethical approval from their own institution to undertake the research.

- **Restricted access** – Metadata are made public, but the full access is granted only to registered users.

- **Closed access** – Metadata is published, but data is inaccessible and there is no process in place to apply for access to it. This type of access is rarely used. Examples include a case when you worked on generating data, but you don't have the right to publish it. Alternatively, it might be classified as military research or another type of sensitive data.

- **Depositor access** – Data can be accessed solely by depositors.

To access data, various types of user agreements exist. The most common type is an open content licence, such as a Creative Commons (CC) or general public licence. You can learn more about different types of licences in the chapter "Open-source, open licensing, scientific programming".

5.4   What exactly is open research data?

According to the European Commission, "open research data refers to data underpinning scientific research results that has no restrictions on its access, enabling anyone to access it."

Making your data open brings a lot of advantages to you, as well as to the whole scientific community. It reduces cost and saves time for the government or private sector when reusing your data for further understanding and knowledge or for other researchers when they don't have to collect the same data again. Many collaborations with other scientists are born this way. By opening your research, you are making it not only more visible but also transparent and you are potentially increasing its impact. Your work can get potential recognition by the general public and in the scientific community. This also encourages scientific enquiry, debate and improvement and validation of research methods. It is possible that others identify errors in your data or methods. This gives you an opportunity to learn from your mistakes and grow as a scientist.

In fact, research integrity is an essential driver of reliable and trustworthy research and scientific discovery. A fundamental principle of the scientific method is reproducibility. The most often used reasons for not meeting the criteria of reproducibility, also known as replicability or repeatability, are: pressure to publish, poor statistical analysis leading to conclusions not supported by results, poor reliability of results, selective reporting and lack of replication within the original environment. Environmental observations and measurements are unique. That is why choosing correct methods and tools and interpretation of results are important in this type of research. On the other side of the spectrum is experimental research which is repeatable by nature. Whether your research is observational or experimental, good data management helps to support your own research integrity, as well as the validity and reproducibility of your research.

It is your responsibility as a researcher along with all other individuals involved in the research process to manage your data well and to make them open. Whether you've already generated data for your research project or are only planning how to do it, it's never too early (or too late) to start considering how to make your data open. Do not forget that you don't own data you collect or generate when working on your research project. Make sure to familiarise yourself with institutional policies and confirm with your PI/supervisor prior to making your research data open. You can learn more about research ethics in the chapter "Open science policy, scientific integrity and ethics".

To make data open is not enough. It must also be FAIR.

5.5 Test your understanding

Loading...

Activities

In a recommend activities section like this one, we will recommend the activities to increase your understanding of the concepts and improve your practical knowledge.

- Create metadata for one of your data files. Metadata standards used in different scientific disciplines, including specifications and various tools, can be found on the Digital Curation Centre website. Other useful schemas with overall structure of metadata are available on the UNC University Libraries website. Pick one that is the closest to your research and give it a go.

- Fill in a DMP template. You can do it online at DMPonline after creating an account or you can download one from the EC Horizon 2020 research and innovation funding programme or the University of Michigan Library.

- Check out a free online MANTRA course from the University of Edinburgh that was created for those who manage digital data as part of their research project.

- Check out the free Data Tree research data management course from NERC.

- Do you have your data management plan? Or have you found during your research that your data could be managed better and with much less effort should you have had the relevant information when starting your project? Share your experience with others on our social media.

Chapter 6

FAIR principles

6.1 What is FAIR?

In a nutshell, FAIR is a set of guiding principles to make data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable.

The FAIR principles were first launched in 2014 at a Lorentz Workshop and officially published in 2016 with the focus on the EU's goal of increasing sharing and reusing of research data. The implementation of the FAIR principles for research data is a requirement imposed by the EU, alongside the EU's request on Open Science & Open Data. It is noteworthy that the FAIR principles are not a standard.

What is in it for you if you make your data FAIR?

The FAIR principles have multiple advantages for researchers. In general, by working in line with the FAIR principles, you can make your research more transparent, collaborative and sustainable and meanwhile facilitate your data management and protect your data's value for future use.

More specifically, you can expect the following by working with the FAIR principles:

- Greater impact and visibility of your research.
- Opportunities for new research collaborations.
- More credit for yourself as a researcher.
- A more efficient data management plan.
- Possibilities for future research.

What is in it for science if you make your data FAIR?

The FAIR principles also bring great benefits to the research community and thereby a fulfilling sense of community commitment to you as a researcher. FAIR principles:

- Enhance scientific enquiry and debate.

- Enable innovation and new data use.

- Increase the efficiency of research due to reusability and replication studies.

- Provide a valuable resource for education and training.

- Encourage the improvement and validation of research methods.

- Enable scrutiny of research results.

- Facilitate transparency and accountability.

6.2   Key concepts to start with if you want to FAIRify your data

PID (persistent identifier)

A PID is a long-lasting reference to a document, a file, a web page or another object. It is usually used for digital objects that are accessible over the Internet but can also be used for physical objects. For example, the PID for a book can be its ISBN (International Standard Book Number). The use of a PID can effectively slow or prevent the damage of "link rot" in citations, which means that the cited URLs "go dead" because the contents are removed for different reasons.

You can encounter all kinds of PIDs in your research work. Here are two of the most frequently used types:

1. **DOI (digital object identifier)**

The use of DOI is to identify academic and professional information, such as research articles, reports, datasets, publications – and in some cases government documents and commercial videos.

Archiving your data with a data DOI as the PID will allow you to be compliant with the FAIR principles and enhance the impact of your research through increased visibility, leading to more citations.

You can read more about DOIs on the official website of the International DOI Foundation (IDF).

2. **ORCID (open researcher and contributor ID)**

How can you find the work of one specific researcher among all the baffling names? ORCID might be your answer. ORCID provides a persistent identity for humans, so that a particular author's contributions to the literature or publications in the humanities can be easily and clearly recognized.

Metadata

In short, you can define Metadata as "data about data".

There are multiple categories of metadata with different definitions, while the following three are the most relevant to the FAIR principles.

Descriptive metadata are data that allow people to discover and identify them through the context or content, including title, author, abstract, keywords, etc.

Structural metadata are data about the project's internal structure and relationships to other objects, including the unit of analysis, data collection method, sampling procedure, etc.

Administrative metadata are data that are relevant for managing the project, including provenance, licence, creation date, file type, etc.

Metadata are not set out from the beginning and forgotten about. Instead, they are subject to changes and updates. Remember to add or modify your metadata continuously throughout the project.

Metadata can help you to play better with the FAIR principles, because metadata are machine-readable and, especially when they have a PID, search engines can easily find them.

6.3   Let's FAIR up!

The principles

The FAIR principles are quite straightforward. Below are the guidelines and you can read about the details for each at the FAIR principles website.

1. **Findable**

F1. Metadata and data are assigned a globally unique and persistent identifier.

F2. Data are described with rich metadata (defined by R1 below).

F3. Metadata clearly and explicitly include the identifier of the data it describes.

F4. Metadata and data are registered or indexed in a searchable resource.

2. **Accessible**

A1. Metadata and data are retrievable by their identifier using a standardised communications protocol:

   A1.1. The protocol is open, free and universally implementable.

   A1.2. The protocol allows for an authentication and authorization procedure, where necessary.

A2. Metadata are accessible, even when the data are no longer available.

3. **Interoperable**

I1. Metadata and data use a formal, accessible, shared and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles:

   Ontologies

   Vocabularies

   Taxonomies

I3. Metadata and data include qualified references to other (meta)data.

4. **Reusable**

R1. Metadata and data are richly described with a plurality of accurate and relevant attributes:

   R1.1. Metadata and data are released with a clear and accessible data usage licence.

   R1.2. Metadata and data are associated with detailed provenance.

   R1.3. Metadata and data meet domain-relevant community standards.

Step by step

There are six FAIRification practices you can do to make your data FAIR.

   Documentation

   File formats

   Metadata

   Access to data

   PID (persistent identifiers)

   Data licences

Documentation

Documentation of data usually happens on two levels:

1. Data-level documentation. At this level you should include information such as data type, data processing procedures, structure of the data, e.g., questions, variables, concepts, etc.
2. Project-level (or study-level) documentation. At this level you should include information such as when, how and why the data were generated and by whom, how the data were processed, what quality assurance measures have been used, etc.

It is noteworthy that the lists are not exhaustive – other information or data files are often included at both levels.

When it comes to publishing and reserving data, FAIR documentation enables you as a researcher to show how the data was generated and for what purpose by including information such as the following:

- Methodology descriptions.

- Codebooks.

- Questionnaires.

- Scripts like do-files editors (STATA).

- Laboratory notebooks and experimental protocols.

- Software syntax and output files.

- Database schemes.

- Provenance information about secondary data.

- Finalised data management plan.

Publishing the documentation together with your data in a repository will boost the re-usability of your data and the likelihood of your data being cited – thus more FAIR data.

File formats

Different file formats have different characteristics and properties. Some can have limitations. It is a good idea to decide the purpose of a file first – for example, data collection/processing/analysis, reuse, or preservation – as it helps to determine which format to use. Sometimes it can be handy to keep some data files in multiple formats.

When it comes to publishing and reserving data, you have to consider whether the file formats used for data collection, processing and analysis are also appropriate formats for long-term preservation. Furthermore, in the spirit of the FAIR principles, choose the right file format for publishing and preserving so that you and others can access and use the data later.

Here are some examples of preferred FAIR file formats for preservation:

- Containers: TAR, GZIP, ZIP

- Databases: XML, CSV, JSON

- Geospatial: SHP, DBF, GeoTIFF, NetCDF

- Video: MPEG, AVI, MXF, MKV

- Sounds: WAVE, AIFF, MP3, MXF, FLAC

- Statistics: DTA, POR, SAS, SAV

- Images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP, SVG

- Tabular data: CSV, TXT

- Text: XML, PDF/A, HTML, JSON, TXT, RTF

- Web archive: WARC

Metadata

Earlier in this chapter, we learned about the concepts and categories of metadata, which play an important role in making your data FAIR. Remember to add metadata continuously to your research data, not just at the beginning or at the end of your project. You can read more about metadata standards and ontologies at Dublin Core and the RDA Metadata Standards Directory Working Group. Don't forget that your metadata must have a findable PID (persistent identifier) that is typically assigned when a digital resource is placed in a data repository.

Access to data

Always consider the following before you make your data accessible:

- Who are the data available for and under which conditions?

- How are the data backed up?

- How is the above documented?

- How may the Intellectual Property Rights (IPR) agreements restrict access to the data sets both during the collection and after finalising the project?

- Do you and your collaborators agree on the 4 points above and the standard procedures and documents?

It might sound a little surprising that sensitive data, which include, for example, personal or confidential information, can also be FAIR without being open. Common practice is to anonymise (change to impersonal ID's) or de-identify (remove ID's) the data. However, this often comes with some limitations. For

example, old, de-identified data cannot be added to new data after a certain period of time, which limits the reusability of the data.

PID (persistent identifiers)

Previously in this chapter, we have gained an understanding about PID (persistent identifiers). But how can you get a PID for your data and metadata? You can start with finding a repository that will provide a PID.

- You may find something interesting and suitable for you on the list of repositories recommended by the European Research Council.

- You can visit Re3data, which is a global registry of research data repositories from various academic disciplines.

- FAIRsharing allows you to discover databases grouped by domain, species or organisation.

- You can also check whether your institution has a local repository that can provide a PID for research data stored at their own local repository.

And there is a lot more if you still want to browse for other repositories, such as OpenAIRE, Figshare, ROAR, etc.

Data licences

A data licence is a legal arrangement between the creator of the data and the end-user/the place for data depositing, which specifies what users can do with the data. The most commonly used data licences are the suite of Creative Commons (CC) copyright licences, which concern reusability of the data and are irrevocable. Another widely known licence is Copyright. You can learn more about licences in the chapter "Open science policy, scientific integrity and ethics".

6.4   Test your understanding

Loading...

Activities

In a recommend activities section like this one, we will recommend the activities to increase your understanding of the concepts and improve your practical knowledge.

- Find an online dataset and investigate how FAIR the dataset is:

    - Findable – Do PID's exist? Are metadata searchable?
    - Accessible – Are metadata and data retrievable?
    - Interoperable – Using open file format? API?

- – Reusable – Can you find the provenance, licence and description of data?

- Try to answer the following questions:

  - – Will you consider making your data FAIR? Why/why not?
  - – What do you think the advantages/disadvantages could be?

- Share your experience with FAIR principles in your research with others on our social media.

Chapter 7

Data repositories and data centres

Long-term preservation and sharing of data are a part of the data lifecycle and should be a part of your data management plan that you learned about in the "Research data" chapter. What's more, it can be one of the requirements of receiving funding. Therefore, storing research data of long-term value in data repositories or data centres should be an important part of your planning process.

7.1 Data repositories

A data repository is a set of data that has been isolated for the purpose of data reporting and analysis. The data repository is a multi-database infrastructure that collects, manages and stores data sets for data analysis, sharing and reporting. Grouping data together enables easier and faster data analysis and using data repositories is a part of data management. Data can also be stored and archived in data repositories.

Examples of types of data repositories include the following:

- **Data warehouse** – A data warehouse is a system that collects data from several sources into a single, central, consistent data storage system for use in data analysis, data mining, artificial intelligence and machine learning. Data warehousing helps in improving data quality by centralising data from various sources, including transactional systems, operational databases and flat files. The file is then cleansed and standardised and duplicates are removed to establish a single source for the file.

- **Data lake** – A data lake is a data warehouse without predefined schemas. As a result, it supports a wider range of analytics than a traditional data warehouse. The main distinction between a data lake and a data warehouse is that data lakes contain large amounts of unstructured data whereas the data in a data warehouse is structured.

- **Data mart** – A data mart is a simplified data warehouse which focuses on a single subject. It reduces the time and resources required to go through more complex data in a data warehouse or manually aggregating data from several sources. By using a data mart, teams can access data and obtain insights faster than using a data warehouse.

- **Metadata repositories** – Metadata provides fundamental information about data, making it easier to discover and operate with specific types of data. A metadata repository is a software which maintains descriptive data about the data model that is used to store and exchange metadata. Diagrams and text are combined in metadata repositories, allowing for metadata integration and updating. A metadata repository can also serve as the foundation for a data warehouse, among other things.

- **Data cubes** – A data cube is a representation of the precise information that has to be extracted from a large quantity of complex data. A data cube is a structure designed to handle analytical queries. They contain metrics that have one or more regularly used dimensions. The metrics are precomputed, which means that a database job takes raw data, does computations and produces a new table to hold the results so that they can be queried instead of having to query and compute them from scratch each time. When data is aggregated depending on particular factors, numerous tables are layered on top of each other, creating a cube-like result. Data cubes group together various important data to allow for more flexible analysis, such as spotting trends. When data is organised into cubes, queries may be run significantly faster than if the tables were all kept separately.

- **Operational data stores** – An operational data store is used to keep comprehensive transactional data from several operating systems. It can function as a stand-alone operational system or as a temporary staging location for data until it is cleansed, processed and put into a data warehouse. An ODS receives data from other systems on a continual basis, either through real-time data replication or batch extract-transform-load processes. In most cases, data is kept in a denormalized manner. Operational data stores are useful for querying small datasets in real-time or near-real-time (Kumar, 2019).

Re3data – research data repositories registry

A research data repository is the best approach to publishing and exchanging research data. A repository is an online database that allows long-term preservation of research data while also assisting others in finding it. A repository will provide a DOI for each submitted object and provide a web page that describes what it is, how to cite it and how many times other researchers have cited or downloaded it, in addition to archiving research data.

The Re3data registry of research data repositories is an open scientific platform that provides an insight into the existing international research data repositories for academics, funding agencies, libraries and publications. The Humboldt University in Berlin, the German Research Centre for Geosciences (GFZ), the Karlsruhe Institute of Technology (KIT) and Purdue University all contribute to the register. By using a key-term search, the Re3data catalogue can be used to search for datasets. Results can also be filtered to show different fields such as social sciences, economics, anthropology, geography and the humanities. Also, the three browse functions – subject, content type and country – can be used to locate datasets. Data can be found by topic area using an interactive chart of disciplines and subdisciplines. Furthermore, data regarding the source repository are stated in the series metadata.

Some examples of popular data repositories for storing research data are figshare, Mendeley Data, Dryad, Zenodo and Open Science Framework.

7.2 Data centres

While a data repository is a type of data library or data archive, a data centre also provides data asset management and services for the entire organisation through a series of platforms, tools, processes and specifications. A data centre refers to a location where data and applications are stored. Data centres make deposited data available to other scientists, commercial enterprises, government bodies, educational institutions and the general public. The architecture of a data centre is built on a network of computer and storage resources that allow shared applications and data to be transmitted. Data centres' architecture and functionalities have been greatly improved over time. The traditional on-premises physical server with virtual networks that support applications and workloads across groups of physical infrastructure has evolved into a multi-cloud system. As a result, data is now available and networked over numerous data centres, the edge and public and private clouds. Over time, data centres have increased their capacity to communicate with different locations both on-premises and in the cloud.

Components of a data centre

The primary components of a data centre are the computing unit, storage unit and network unit. The components of a data centre include computers, servers, routers or switches, a firewall or biometric security system, storage systems such as storage area networks or backup/tape storage, data centre management software/applications, power and cooling devices such as air conditioners or generators, physical server racks and chassis cables and an internet backbone.

Data centre services

Data centres can help you to develop a DMP. Data centres offer services such as data storage, backup and recovery, data management and transfer for data of long-term value. Data can be made accessible openly or subject to restrictions and data centres often provide a searchable catalogue of data stored therein. Finally, data centres can issue DOIs.

How to find the right data centre for research?

There are some critical factors to be considered in the choice of a data centre to host research data. After finding a data centre that specialises in a particular branch of scientific research, other considerations should include location, reliability, security, network service capacity, flexibility and scalability, emergency backup and reputation. If you are not sure you chose the right data centre, contact them. They'll be able to confirm or provide advice about better alternatives.

How to deposit data in a data centre?

The process of depositing data starts with preparation of the data, validating the data (performing quality checks) and making it available in a form acceptable to the data centre. Before uploading the data to the data centre server, the guidelines stipulated by the data centre must be read and understood. Also, the owner of the data must be clearly defined and a brief description of the data should be provided. While open access of data is preferred, the data owner is expected to specify the type of access they wish to grant to their data. You can learn more about types of open access in the chapter "Open data and open access". Data are then uploaded and the upload is confirmed.

Because data cannot be modified after the upload, data must be in their final version before being deposited. The DOI for publication is also required when depositing data that contain datasets which support the publication, but some data centres can issue a DOI. You can learn more about DOI in the chapter "FAIR principles". The owner of the data is expected to indicate the type of

access they wish to grant to their data. When a data set should have more than one type of open access, data must be stored separately in separate sections. Any scientific unit or research group's data must also be validated before being deposited in a data centre.

Importance of PIDs

A persistent identifier (PID) is a permanent and unique digital identification that helps users to locate and reuse digital content. PIDs are numbers that are used to identify digital items such as files, documents and web pages. Even if an organisation's web address changes, the PID will continue to function and ensures that a link to a digital item remains functional even if the object's location or web URL (Uniform Resource Locator) changes. Various options for a PID include Digital Object Identifiers (DOIs), Persistent Uniform Resource Locators (PURLs) and Archival Resource Keys (ARKs). Handle and OpenURL are examples of PID systems. Each system has its own set of characteristics, strengths and flaws.

Every dataset that is deposited and made accessible through a data centre will have a DOI, which is a form of PID. This ensures that the research is correctly cited, as well as indicating which version of the data was used. It is frequently a prerequisite to publish material based on these datasets in a publication where the research data are accessible and have a PID.

Trends and future development

Due to the high energy consumption of data centres, there is a current trend to shift to renewable sources of energy. Renewable energy sourcing is standard, most data operators pay a premium to their suppliers for a 100% supply of certified renewable energy. Direct partnerships with renewable energy plants are becoming more popular globally. This demand for renewable energy will give the renewable industry much-needed investment, certainty and it may even minimise the need for REFIT (renewable energy feed in tariff) programmes for renewable energy.

7.3 Test your understanding

Loading...

Activities

In a recommend activities section like this one, we will recommend the activities to increase your understanding of the concepts and improve your practical knowledge.

- Check out the Digital Repository of Ireland, the UK Polar Data Centre and the National Centre for Biotechnology Information. How would you deposit data in each of them?

- Identify a data repository/data centre suitable for depositing your research data. Add it to your data management plan.

- Did you find the previous two activities easy? Or were they a bit of a challenge? Share your experience and tips on data repositories/centres suitable for a specific type of data or research with others on our social media.

Chapter 8

Open science policies, scientific integrity and ethics

8.1 Open science policies

Open science policies are a set of guidelines and protocols developed and implemented to promote open science principles. These policies can be mandatory or voluntary in nature and are often put in place by government agencies, funders, research institutions, etc. Often, these policies provide a framework for making research data accessible to the public long term and to evaluate its subsequent impact. If you are interested in adopting open science practices in your research, the first step is to understand your organisation and relevant stakeholders, like the funding body, their current policies and how they can affect your work. If you would like to know more about organisations and their policies, check out the European Commission's Ethics and Data Protection guidelines.

The most common factors addressed by policymakers to increase effectiveness are:

- Provision to review and identify the nature of the data and suitability to release it for public access.

- Provision to restrict or anonymize personal data.

- Provision of data centres and repositories to store and publish the data.

- Provision of roadmaps, tools, checklists, etc. for open-access publication.

- Outline of the intellectual property rights, copyrights, data licensing and reuse rights.

- Outlining the data publishing standards in line with open standards.

- Provision to monitor, assess and review the policy.

- Data anonymization.

You can learn more about storing public data in the chapter "Data repositories and data centres", about open access publishing in the chapter "Open data and open access" and about data licensing in the chapter "Open-source, open licensing, scientific programming".

8.2   Scientific integrity and ethics

In order to maintain the integrity, value and benefits of open science research there is a need to adhere to ethical principles across disciplines which include societal issues such as non-discrimination, privacy and confidentiality as well as professional conduct. Due to variations of these across countries, local laws and cultural norms must be considered and adhered to. It is useful to consider that these principles could evolve due to factors which include historical events, scientific advancements and behavioural norm changes.

While there are national and disciplinary differences in the way research is organised and conducted, there are also general principles and professional responsibilities that are fundamental to the integrity of research as indicated in the Singapore Statement on Research Integrity, published in 2010.

General principles:

- **Honesty** in all aspects of research.

- **Accountability** in the conduct of research.

- **Professional courtesy and fairness** in working with others.

- **Good stewardship** of research on behalf of others.

Professional responsibilities:

- **Integrity** – Researchers should take responsibility for the trustworthiness of their research.

- **Adherence to regulations** – Researchers should be aware of and adhere to regulations and policies related to research.

- **Research methods** – Researchers should employ appropriate research methods, base conclusions on critical analysis of the evidence and report findings and interpretations fully and objectively.

- **Research records** – Researchers should keep clear, accurate records of all research in ways that will allow verification and replication of their work by others.

- **Research findings** – Researchers should share data and findings openly and promptly as soon as they have had an opportunity to establish priority and ownership claims.

- **Authorship** – Researchers should take responsibility for their contributions to all publications, funding applications, reports and other representations of their research. Lists of authors should include all those and only those who meet applicable authorship criteria.

- **Publication acknowledgement** – Researchers should acknowledge in publications the names and roles of those who made significant contributions to the research, including writers, funders, sponsors, and others, but do not meet authorship criteria.

- **Peer review** – Researchers should provide fair, prompt and rigorous evaluations and respect confidentiality when reviewing others' work.

- **Conflict of interest** – Researchers should disclose financial and other conflicts of interest that could compromise the trustworthiness of their work in research proposals, publications and public communications as well as in all review activities.

- **Public communication** – Researchers should limit professional comments to their recognized expertise when engaged in public discussions about the application and importance of research findings and clearly distinguish professional comments from opinions based on personal views.

- **Reporting irresponsible research practices** – Researchers should report to the appropriate authorities any suspected research misconduct, including fabrication, falsification or plagiarism, and other irresponsible research practices that undermine the trustworthiness of research, such as carelessness, improperly listing authors, failing to report conflicting data, or the use of misleading analytical methods.

- **Responding to irresponsible research practices** – Research institutions, as well as journals, professional organisations and agencies that have commitments to research, should have procedures for responding to allegations of misconduct and other irresponsible research practices and for protecting those who report such behaviour in good faith. When misconduct or other irresponsible research practice is confirmed, appropriate actions should be taken promptly, including correcting the research record.

- **Research environments** – Research institutions should create and sustain environments that encourage integrity through education, clear policies, and reasonable standards for advancement, while fostering work environments that support research integrity.

- **Societal considerations** – Researchers and research institutions should recognize that they have an ethical obligation to weigh societal benefits against risks inherent in their work.

8.3 Research misconduct

Violation of research conduct or research ethics when proposing, performing or reviewing research or reporting research results is known as research misconduct. Under the umbrella of research misconduct, we can find fabrication, falsification and plagiarism. Other irresponsible research practices that can undermine the trustworthiness of research but that are seen as having less impact on the research process are often referred to as questionable research practices. Their definition and categorisation are country-dependent. According to the European Commission's Report on Responsible Open Science in addition to upholding high quality research through transparent and reproducible results, there is also a need to deal with challenges which result from malpractices such as fake science, biassed assessment and predatory journals.

To keep the research code of conduct up to date and in line with emerging changes it is frequently reviewed. Mechanisms to conduct ethical research have been established, including the peer review process, policy governance structures and education and training of researchers. No one is implying that you would intentionally commit research misconduct. However, it is good to know what to watch out for. A good place to start reading about it is the Policy Brief on Research Integrity by the European Commission, the OECD's Best Practices for Ensuring Scientific Integrity and Preventing Misconduct and policies at your institution.

8.4 Research with human subjects

If your research involves humans as research participants, it must firstly be approved by an independent review body. Research Ethics Committees or Institutional Review Boards provide you with helpful advice and will answer any questions or doubts you might have about your research. Basic guidelines are provided:

- In the Nuremberg Code (1947).

- By the Council of Europe (CoE).

- In the Declaration of Helsinki (1964).

- In the Convention for the Protection of Human Rights and Dignity of the Human Being (1997).

If your research involves human participants or if you are dealing with personal data, appropriate tools should be used and the attention to the data protection and privacy laws is of utmost importance. Researchers should:

- Have informed consent from participants.

- Ensure confidentiality by protecting participants' anonymity.

- Avoid any practices considered to be deceptive.

- Ensure participants have the right to withdraw if required.

To receive fully informed consent from human subjects is mandatory. Informed consent must be:

- **Clear** – To achieve this use plain language.

- **Concise** – Don't overwhelm the potential participant with unnecessary information.

- **Continuous** – As the research progresses participants must confirm they wish to continue.

Data protection law updates should be noted with regard to new collection, processing, storage, updating and enforcement methods. For example, the new EU General Data Protection Regulation (GDPR). Different countries have different data protection authorities, who provide guidance on their specific laws.

8.5 Research with animal subjects

As animals deserve respect but cannot argue for their own protection and rights, they must be protected if used in your research.

Generally accepted guidelines (Russell & Burch, 1959) for research with animal subjects are:

- **Replacement** of use of animals with other alternatives wherever possible.

- **Reduction** in the numbers of animals used.

- **Refinement** of research techniques to reduce discomfort, distress or pain for the animals being studied.

Legislation relevant to research on animals varies greatly between countries, even within the EU. Analogous to research with human participants, an independent committee oversees research with animals. Committees will have different names in different countries.

8.6   Test your understanding

Loading...

Activities

In a recommend activities section like this one, we will recommend the activities to increase your understanding of the concepts and improve your practical knowledge.

- With all the online resources available today we have access to the work of other researchers, now more than ever. On the other hand, we also have many tools at hand to avoid plagiarism, even if unintentional. Check if your university offers a specific plagiarism checker or, alternatively, try Plagiarism Checker or Plagiarism Detector.

- Are you wondering about the research misconduct investigation procedure? You can find more information on, or study the procedure for, the investigation of misconduct in research at the UK Research Integrity Office website or the UCD website.

- Have a look at some articles that are around a decade old discussing research misconduct. Or have you noticed such an article recently? Do you think a lot has changed in the last 10 years?

- Have you learned anything new or shocking? Can you imagine a situation when you could be under pressure to commit research misconduct or irresponsible research practices? Share your thoughts with others on our social media.

- Check OSF – a free, open source web application that connects and supports the research workflow, enabling scientists to increase the efficiency and effectiveness of their research. Researchers use OSF to collaborate, document, archive, share, and register research projects, materials, and data.

Chapter 9

Communication and dissemination

Now that you have gathered a background to open science, you are probably wondering what you should do next? Where do you even begin? As researchers, we are often encouraged to focus on academic communication such as journal articles and conference presentations. While this is the goal, there are so many other ways in which you can share your research, even before you get to the point of sharing your results. We are taught to use a language style that is often inaccessible to the majority. This chapter is going to take you through how open publications work, while also opening your eyes to other methods of communicating and disseminating your research in a more open and accessible way.

9.1   What exactly is scientific communication?

We often hear the terms dissemination and communication, but what do these terms mean? Wilson *et al.* (2010) defined dissemination as "a planned process that involves consideration of target audiences and the settings in which research findings are to be received and, where appropriate, communicating and interacting with wider policy and health service audiences in ways that will facilitate research uptake in decision-making processes and practice."

So, when we talk about **dissemination**, we are talking about the targeted activities in which you share information with others in the science community, stakeholders and even policy makers. It is focused on industry-specific knowledge exchange. This usually comes in the form of scientific journal publications or conference presentations that are often presented in a scientific language which is usually inaccessible to the general population.

Communication, on the other hand, is more focused on the public facing sharing of information. This is, by nature, the more open form of sharing information. The language is written in a lay format which focuses less on the scientific representation of results and more on the interpretation and impact of your results. This is often done through public events, newsletters, social media, etc.

For more information on the differences between dissemination and communication practices check out LeitaT Project's blog post.

9.2   Open science publishing

While dissemination activities are often targeted, there are some publishing routes you can take to make your scientific publications more open. Publishing in open science allows you to share your full research journey, from conception of the idea and the research protocol to the results and impact of your research. The next few paragraphs will take you through different ways of disseminating your research through publications.

- **Pre-registration**

To pre-register a study means that you specify the research plan and hypotheses prior to the data collection phase and submit this information to a public repository. This allows peers to review your plans, make comments and provide feedback before you begin the data collection period. Pre-registration also helps with some of the issues around publication biases. Examples of this are things such as not publishing negative results or hypothesising after results are known (HARKing). When you pre-register your study, the journal conditionally accepts your future paper on this study regardless of the results, even if the results are not as you planned. While pre-registration is not a necessity for open science, it does allow for your research to be more accessible from the beginning.

If you're looking to pre-register your study, a good place to start is the Open Science Framework. This is an easy-to-follow method for pre-registering your research in a multidisciplinary repository.

- **Registered reports**

The next step in the open science publishing route is registered reports. This is where you go a step further from pre-registration and you submit your plan or protocol for your study to a journal. This report is focused on your research questions and methodology, but similarly to pre-registration, the journal you submit your registered report will agree to publish the results of your study if

the reviewers believe that the plan for the study is clear and reaches a high scientific standard. These types of publications are most popular in the sciences in which a bias exists, for example in psychology, but are useful in all disciplines.

For more information on registered reports and to find a journal that suits your discipline, check out the Centre for Open Science's website.

- **Open access journals**

Now that you've gathered your data and analysed it, the time has come to share this with the public.

Article Processing Charges – Much like publishing articles in the traditional sense, article processing charges (APC) exist in open access journals. The Directory of Open Access Journals provides a worldwide overview of what journals require APCs and, if they do, how much it will cost. Not all journals charge costs. And very possibly, your institution has a contract with many open access journals where you can publish for free.

You can also learn about types of open access publishing in the chapter "Open data and open access".

9.3 How to communicate your research

Now let's focus on how to communicate your research with the public. There are a few easy methods of doing it and making it accessible to the general reader. When using these methods, remember to avoid all jargon and statistics. The focus here is on what the results of your research mean and how they can be used. **Lay summaries**

A good place to start when it comes to communication are lay summaries. A lay summary is a short summary avoiding academic jargon which describes your research. A good lay summary would be accessible to those who are outside of their specialty but also those in the general public. A good way to judge your summary is to ask yourself – if I walked into a high school and described my research, would the students understand it? If yes, then well done. If not, then try again. **Social Media**

Social media is a handy tool to share your research, if used correctly. These platforms allow you to showcase your research activities quickly and frequently for free. Social media platforms also offer this opportunity to see the impact your research is having in real time through live engagement metrics. This live engagement element gives you the chance to see who is interested in the research topic, who is sharing your research and who is talking about the topic. It can be time consuming to devote your efforts to managing social media accounts, especially if you manage more than one, but it can be worth it when people engage and ask questions about your research.

Social media is not new to academia. In fact, academia-specific platforms are being used in the form of ResearchGate and Academia. Despite this, many researchers chose to use the more "social" platforms such as Twitter and LinkedIn to share research and join conversations. More recently, Instagram and TikTok are being used by researchers to share their research through visual means. A good example of this is @science.sam, a well-known Canadian science communicator who has made a career out of making science accessible and approachable to a general audience. Dr. Samantha Yammine uses multiple social platforms and media forms to showcase general science and her own specific research. You can check out her other social media platforms to see scientific communication in action.

Check out this blog on the Academic Designer for a concise outline of all the different social media platforms available to you. Use this information to decide what platforms work for you and your research. Two questions to keep in mind when setting up a social media account for your research are "who is my target audience?" and "what platform will they use?". If your research will benefit teenagers, then creating a Facebook page will not work as that's not the platform teenagers use. **Blog posts**

Blogs are a great way to share your research results, as well as your research journey. A lot of science communication is about sharing a story. Blogs open up this opportunity to share your research through stories in a reflexive manner. Not only does it offer you the opportunity to share the research, but you can also share the process you went through in undertaking the research. This offers a learning opportunity for others who wish to use similar methods to you.

The main barrier to blog posts is the platform you use. Setting up and managing an individual website to host a blog can be a time-consuming activity when you are in the middle of research activities. Many people do not want to host their own website, so they use institution or organisation sites instead. An example of an organisation's blog would be the PLOS SciComm blog. The website hosts two types of blogs, those written by a team of staff members who are trained science communicators and independent blogs. Independent blogs are suggested by researchers on topics they write about as a guest on the website. For more information on this, check out the about section of this website.

Institutional blog posts are those which are hosted by the organisation supporting your research. Some institutions have blogs on the home website, while many research groups have their own individual blogs specific to their areas of interest. It is best to ask around and see if there is a blog already up and running. If not, make a suggestion to the communications team.

For more information and examples on science communication blogs check out this post by Dr. Miguel Balbin on Animate Your Science website. **Podcasts**

Podcasts have been taking over the world in the last few years, so it's no surprise to see the science community start to take part. Podcasts are actually the audible form of a blog post. They are easily accessed on devices which connect

to the Internet, allowing the general public the opportunity to listen in and learn. Some can be quite informal and conversational, while others are more organised and focused on a specific topic with a strict structure and questions. Similarly to blog posts, a major con to podcasts is the time-consuming element. Podcasts can be a lot of work, especially for one person. There are many steps, from idea generation to the actual recording of the episode, editing, uploading and the many other steps in between. Putting together a comprehensive podcast series takes patience and skill. A good first step if you are interested in podcasting is to be a guest for an episode. This allows you to get some experience and learn before you delve into the process.

Some relevant examples of podcast series which bring in multiple guests to share their own expertise and experience are ReproducibiliTEA and ORION Open Science. These offer you more learning on the topic of open science. **Events**

One final activity that encourages public engagement is public-facing events. These may be pre-existing events hosted in the community where you give a talk about your research or maybe an evening organised by your research group to showcase the work you are doing and the impact it may have on the local community. These public-facing activities don't need to be large or difficult to organise, it could be as simple as giving a talk in a local school or joining an already established event in an even more relaxed environment like a pub. Yes, the local bar. Check out the Pint of Science website to see where your local event is and to sign up. Simple activities like these help you to gain more confidence and become a well-rounded science communicator.

9.4   Accessibility

We have spoken about making your research more accessible to the general public through lay language, but let's not forget making your research more accessible to those with accessibility issues. This is something of particular interest when we speak about communication. When creating social media posts there are a few considerations dependent on the mode of media you choose:

- **Text** – Is it easily transcribed into another language?

- **Images** – Can you provide alt-text to describe what is in the image for those who are colour blind or who may have difficulty seeing full images?

- **Videos** – Have you included subtitles for those who are hard of hearing?

Take a look at some of the free resources on Ability Net to learn more about how you can make your communications even more accessible to those with extra needs. **Bropen science**

Before we close out the chapter, one concept you should be aware of when going down the open science route of publishing and sharing your research is the term "bropen science". This term came about following a viral tweet in 2017 after a female-led research team had their open research criticised. The term refers to the idea that open science practices, while they aim to be accessible, are still led by the rigid and hostile academic environment, which encouraged the open science movement to be born. While the term "bro" is used, this is not to suggest that this is a male-only agenda. A hostile environment can be created by anyone who is a "bro" and tries to discredit and, in many ways, bully individuals who are sharing the good, the bad and the ugly of their research. The problem does not just lie in the specifics of this event. This is not a debate on open science and gender, but a debate on open science and marginalisation.

When moving forward with your open science journey, the message here is – don't be a bro! If you want to read more into the topic, check out Pownall, Talbot and Henschel's 2021 paper or for a more comprehensive overview of the events that led to the formation of the term, take a look at Craig Harper's post.
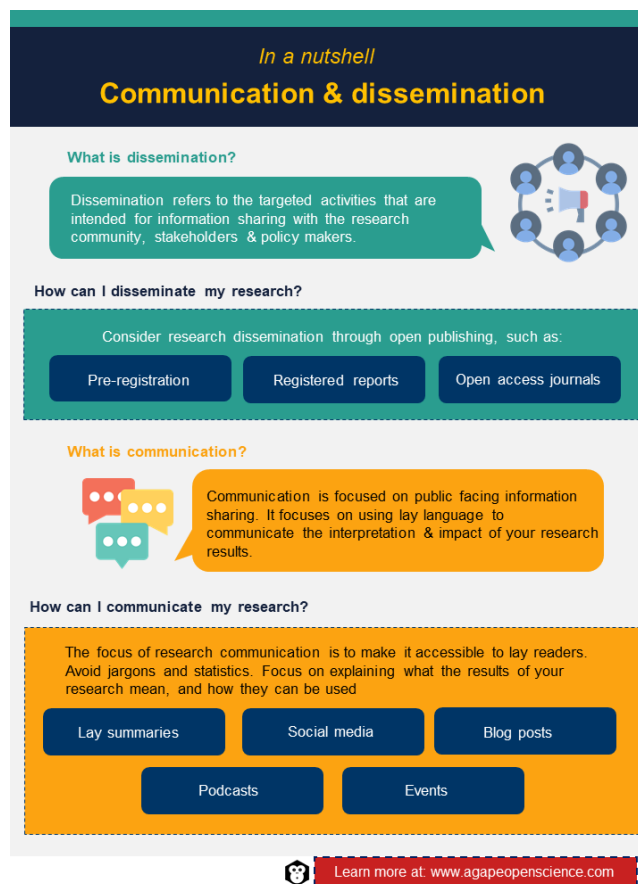
9.5 Test your understanding

Loading...

Activities

In a recommend activities section like this one, we will recommend the activities to increase your understanding of the concepts and improve your practical knowledge.

- Take some time to think about your research and how you want to share the knowledge that comes from it. Speak to your research team about possible dissemination and communication activities and work together to make a plan.

- Pre-register your research. Check out what the standard is in your discipline. Once you've done so, share a link with others on our social media and don't forget to include the discipline of your research. This will help others who are also on this journey.

- Create a lay summary of your research. Describe your research in 30 words or less. Share it with others on our social media.

- Draft up a blog post which could be submitted for publication to your institution's blog. It could even be about your experience and plans for integrating open science into your research. If you're feeling extra brave, send it in. Once published be sure to share the link with others on our social media so we can see your learning in action.

- Why not give designing an accessible infographic a go? Design an infographic explaining your research and share it with others on our social media. Try out free software such as Canva or Piktochart.

Final quiz

You have unlimited attempts to successfully complete the final quiz but only five minutes for each attempt. If you score 90% or higher a certificate of completion will be generated for you and you can download it directly. Click "Start Quiz" when you are ready.

Loading...

Evaluation survey

Thank you for participating in the course. We would like to understand your experience with the Agape course to improve the course. Whether you simply browsed or completed the course, your feedback is valuable to us. Thank you for taking this survey. Please click the link below to take the survey.

Evaluation survey