Istituto di Scienza e Tecnologie
dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche

ISTI Technical Reports

Research workflows and Open Science

Leonardo Candela, ISTI-CNR, Pisa, Italy

Donatella Castelli, ISTI-CNR, Pisa, Italy

Dario Mangione, ISTI-CNR, Pisa, Italy

ISTI-TR-2022/026

Research workflows and Open Science
Candela L., Castelli D., Mangione D.
ISTI-TR-2022/026

The open science paradigm is increasingly praised and encouraged for improving efficiency through deduplication of efforts and for ultimately accelerating scientific discoveries. Such shift towards a collaborative and inclusive scientific process implies an alteration of the traditional research workflows to include the different dimensions that characterise the new paradigm, from open access to new assessment metrics. This systematic study analyses the open science research workflows proposed so far, highlighting (i) their distribution over time, (ii) the various means and approaches used for communicating them, (iii) the terminology used for denominating them, and (iv) the scientific domains where workflows were proposed. Moreover, the workflows were analysed and compared concerning a set of open science aspects deriving from the UNESCO Recommendation on Open Science. Overall, a total of 40 relevant studies were identified and analysed, corresponding to 33 unique workflows. The findings highlight (a) the limited effort spent by the research community to propose and communicate workflows oriented to match open science requirements, and (b) the different nuances of the meaning and understanding of open science and the resulting gap between its theoretical aspects and its practical application to the research processes.

Keywords: Open Science, Open Science Workflow, Survey.

Citation
Candela L., Castelli D., Mangione D., Research workflows and Open Science. ISTI Technical Reports 2022/026. DOI: 10.32079/ISTI-TR-2022/026.

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Area della Ricerca CNR di Pisa
Via G. Moruzzi 1
56124 Pisa Italy
http://www.isti.cnr.it

Technical Report, ISTI-2022-TR/026, 2022
Consiglio Nazionale delle Ricerche - Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

Research Workflows and Open Science

Leonardo Candela, Donatella Castelli, Dario Mangione*

Abstract

*The open science paradigm is increasingly praised and encouraged for improving efficiency through deduplication of efforts and for ultimately accelerating scientific discoveries. Such shift towards a collaborative and inclusive scientific process implies an alteration of the traditional research workflows to include the different dimensions that characterise the new paradigm, from open access to new assessment metrics. This systematic study analyses the literature on open science research workflows and the therein proposed workflows to investigate how these workflows are scientifically communicated and how they respond to the open science needs. With regards to the literature, the study highlights (*i*) the distribution over time, (*ii*) the various means and approaches used for communicating open science workflows, (*iii*) the terminology used for denominating open science workflows, and (*iv*) the scientific domains where open science workflows were proposed. With regard to the workflows, they are analysed and compared concerning a set of open science aspects deriving from the UNESCO Recommendation on Open Science. A total of 40 relevant studies and 33 unique workflows were identified and analysed. The findings highlight (*a*) the limited effort spent by the research community to propose and communicate open science workflows, and (*b*) the different nuances of the meaning and understanding of open science and the resulting gap between its theoretical aspects and its practical application to the research processes.*

Keywords
Open Science — Open Science Workflow — Survey

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Consiglio Nazionale delle Ricerche, Via G. Moruzzi 1, 56124, Pisa, Italy
***Corresponding author**: dario.mangione@isti.cnr.it

Contents

1. Introduction

The open science paradigm is impacting scientific research. Based on a fundamental cultural change enabled and encouraged by new opportunities offered by technology evolution and characterised by different dimensions and application levels, it encompasses various stakeholders and ultimately involves society as a whole [15]. It is inherently collaborative [56, 47], and it thrives on technology-enabled workflows and interdisciplinary research.

By investigating the existing literature, including a broad range of publications such as reports, presentations, and images published until 2022, this systematic study analyses the research workflows that have been proposed with the aim of enabling an open science approach (hereafter referred to as open science research workflows) in order to examine which aspects are connected to the application of open science and, ultimately, how the willingness to implement open science practice is actually impacting researchers behaviour.

This report is organized as follows. Section 2 clarifies the terminology used in this report. Section 3 describes the methodological approach to this study. Section 4 and Section 5 respectively present and discuss the results. Section 6 concludes the report and illustrates future work. Appendix I briefly describes each of the workflows analysed by the report.

2. Study terminology

In this study, we systematically analysed the scientific publications related to workflows proposed or followed by researchers during their scientific activity. Then, based on this analysis, we identified and examined the open science research workflows.

Scientific publications contributing to our corpus are very varied. They range from journal articles to images (cf. Sec. 4.1). In the remainder of the report, we use the term "corpus" to refer to the whole set of publications pertaining research workflows we identified and analysed and "corpus item" to any single constituent of the corpus.

A workflow, which consists in a representation of a series of stages describing a research process, can be viewed at least from two different perspectives: (*i*) the *researcher perspective*, so from the point of view of the actor carrying out a series of actions during a research activity, and (*ii*) the *research object perspective*, which is focused on how a research object flows through a research activity.

While it could be argued that the two perspectives should be distinguished and the two related concepts designated as "workflows" and "life cycles" respectively, we observed that such distinction between the two concepts is not strictly followed in practice, at least in the analysed corpus, and that the two terms are used interchangeably.

Given also the hybrid nature of some of the analysed workflows, since there are cases in which they refer to phases that include the two perspectives at the same time, we decided to use the term '*workflow*' to address both workflows and life cycles.

3. Methodology

The presented survey was conducted following the systematic mapping study (SMS) methodology proposed by Petersen et al. [44] and Kitchenham [25]. As such, this research was structured in four main stages, namely (*i*) definition of the research questions, (*ii*) identification of relevant bibliographic databases and queries formulation, (*iii*) literature review, and (*iv*) results reporting and analysis, which are described in the current and the following sections.

Research questions

The research questions this study aims at answering can be divided into two categories. The first is directed to analyse the corpus on open science workflows to figure out "how" they are scientifically communicated. The latter addresses the workflows themselves to figure out how these respond to the open science needs.

Corpus-related Research Questions

RQ1: What is the temporal distribution characterising the publication of open science workflows?

RQ2: What are the "means" used to publish open science workflows?

RQ3: What are the terms used for naming open science research workflows?

RQ4: What are the scientific domains where the open science workflows originate from?

Workflow-related Research Questions

RQ5: How do the workflows relate to the different facets that characterise open science?

RQ6: Do these workflows imply different nuances of the meaning of open science?

Databases and queries

To conduct the literature search and develop our corpus on open science workflows the following databases were selected: ACM Digital Library, Google Scholar, IEEEXplore, Open Research Europe, OpenAIRE, ScienceDirect, Scopus, Springer, and Web of Science. Ten relevant keywords organised in four groups (based on lifecycle, workflow, method, and protocol respectively) were used:

- "open science lifecycle" OR "open science life cycle" OR "open research lifecycle" OR "open research life cycle";
- "open science workflow" OR "open research workflow";
- "open science method" OR "open science methodology" OR "open research method" OR "open research methodology";
- "open science protocol" OR "open research protocol".

These keywords were used to develop search strings for retrieving publications based on their title, abstract, or keywords, when possible, across the selected databases.

The queries returned 252 results (cf. Tab. 1), further refined to 216 unique entries by enriching and reconciling the DOI-less entries and by removing the duplicates.

Table 1. Retrieved results per database.

Database	Results
ACM Digital Library	1
Google Scholar	127
IEEEXplore	3
Open Research Europe	0
OpenAIRE	70
ScienceDirect	4
Scopus	27
Springer	4
Web of Science	16

The 216 entries were further analysed and reconsidered by using the inclusion and exclusion criteria defined in Tab. 2.

Table 2. Corpus items inclusion and exclusion criteria.

Type	Criterion
Inclusion	The item presents a workflow, or a part of it, and is linked to the open science paradigm; The item is written in English.
Exclusion	The item is not available (e.g., it is published by closed repository); The item is a dataset or a code; The item is actually a pre-print or an image of another item in the corpus;

To reinforce the resulting corpus, the snowballing strategy was exploited, i.e., the references of the corpus items were analysed to eventually identify new items relevant for the study and not retrieved by the above queries.

Ultimately, the corpus for the study resulting from this process consists of 40 items, from which we identified 33 unique workflows. In fact, while there are items presenting more than one workflow there is also the case of many items dedicated to the same workflow (Tab. 3).

A description of each workflow is given in Appendix I.

4. Results

Following the distinction between the two research question groups, the results are presented in the following paragraphs distinguishing between those pertaining to the corpus (Sec. 4.1) and those pertaining to the identified workflows (Sec. 4.2).

4.1 Corpus Features

The analysis of the corpus and its items helps understanding a number of characteristics on the research workflows for open science phenomenon: (*i*) the temporal distributions of the studies documenting and proposing them (RQ1), (*ii*) the means used for publishing (RQ2), (*iii*) the terms used to refer to them (RQ3), and (*iv*) the scientific domains where these workflows were proposed (RQ4).

4.1.1 Temporal distribution

Fig. 1 depicts the distribution of the corpus items along the years. Items are distributed in an interval of nine years, spanning from 2014 to 2022. There is not a relevant pattern in the distribution but the gradually increasing attention given to the topic 'research workflow for Open Science' from 2014 to 2017 and the different averages in the number of publications before and from 2017 onward, 2.3 and 5.5 respectively. 2017 and 2021 are the peaks of the distribution, counting eight and nine publications respectively.

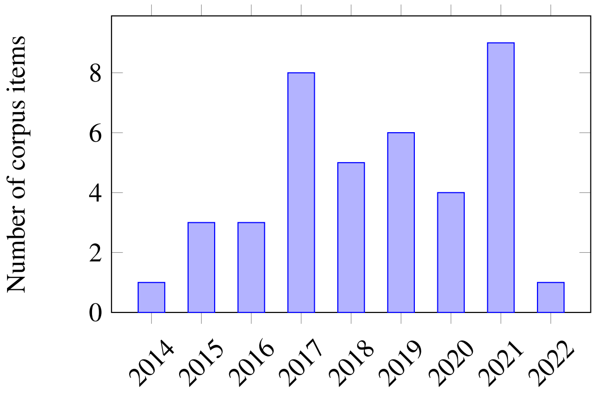


Figure 1. Corpus items yearly distribution

4.1.2 Publishing means

By publication means it is meant the way the research workflow were "published", i.e., scientifically communicated and documented.

Fig. 2 depicts the frequency of the various means. A total of ten diverse means were used: book, book chapter, conference object, conference paper, image, journal article, other literature type, other research product, poster, and preprint. The majority of the corpus items are journal articles (17), followed by conference objects (7) and other literature type resources (5 presentations given at a workshops or at another type of meetings).

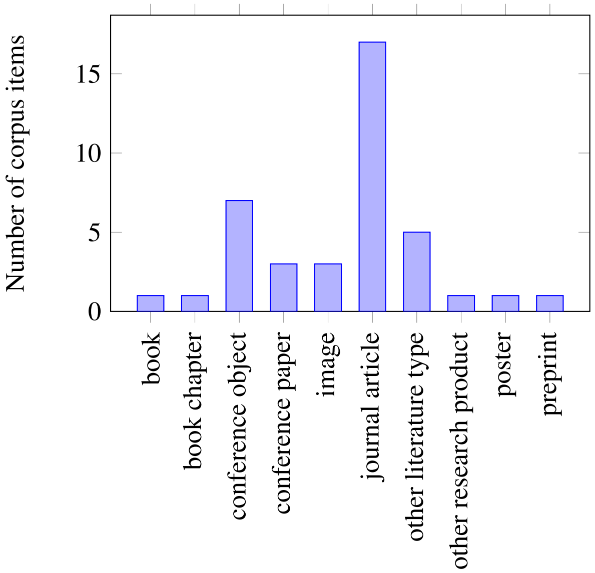


Figure 2. Corpus item typologies

The considerable amount of grey literature items in the corpus reflects how the proposed research workflows are considered a practical issue rather than a theoretical one. This negatively affects the quality of the information provided and leads to poorly documented ones having a limited impact on the community in the large.

4.1.3 Terminology

The terminology used for referring to open science research workflows is not standardised, rather it varies a lot among the items in our corpus.

Two main families were identified, formed around the terms "workflows" and "lifecycles", plus a residual one for the "others".

The workflow-based family of terms stems from 23 corpus items and includes 'collabor* scientif* workflow', 'ohdsi research flow', 'open apc workflow', 'open scienc* paradigm workflow', 'open scienc* research workflow', 'open scienc* workflow', 'open scienc* / open scholarship workflow', 'open workflow', 'open-sci* workflow', 'reproduc* workflow', 'research workflow', 'research workflow cycl*', 'scientif* workflow', and 'workflow for open reproduc* code in scienc*'.

Table 3. Open Science Research Workflows

ID	References	Scientific Domain	Designation	Year
W01	Hripcsak et al. [23]	Medical and health sciences	OHDSI research flow	2021
W02	Harald et al. [21]	Natural sciences	Open Science Workflow	2017
W03	Sarretta [50]; Minelli et al. [36]	Cross-domain	Research data lifecycle	2018
W04	Firth et al. [17]	Engineering and technology	Open Science approach	2018
W05	Firth et al. [16]	Engineering and technology	Open Science Workflow	2018
W06	Linehan et al. [33]	Social sciences	Open research lifecycle	2020
W07	Morissette et al. [38]	Social sciences	Open science / open scholarship workflow	2021
W08	Ignat [24]	Cross-domain	Scientific phases	2019
W09	Puren and Riondet [46]	Cross-domain	Research data lifecycle	2016
W10	Open Science and Research Initiative [43], Muftic [40]	Cross-domain	Research process	2014
W11	Lahti et al. [32]	Humanities and the arts	Reproducible Workflows	2015
W12	Hampton et al. [20]	Natural sciences	Open science workflow	2015
W13	Hampton et al. [20]	Natural sciences	Open science workflow	2015
W14	Hampton et al. [20]	Natural sciences	Open science workflow	2015
W15	Teplitzky [52]	Natural sciences	Research workflow cycles	2019
W16	Corker [13]	Social sciences	Open Science Workflow	2021
W17	Ayris and Ignat [5]	Cross-domain	Research cycle	2017
W18	Klenk et al. [26]	Medical and health sciences	Open science paradigm	2019
W19	Van Lissa et al. [55]	Cross-domain	Workflow for open reproducible code in science	2021
W20	Wandl-Vogt et al. [57]	Cross-domain	Open workflow	2017
W21	Bastille et al. [6]	Natural sciences	Collaborative scientific workflow	2021
W22	Pieper [45]	Natural sciences	Open APC workflow	2015
W23	Beck et al. [9]	Natural sciences	Open science workflow	2021
W24	Assante et al. [4]	Cross-domain	open science workflow	2019
W25	Grigorov et al. [19], Engineering National Academies of Sciences [14]	Cross-domain	Open Research Lifecycle	2016
W26	Gownaris et al. [18]	Cross-domain	Scientific life cycle	2022
W27	Kramer and Bosman [28], Labastida i Juan [31], Bosman and Kramer [10], Kramer and Bosman [29], Kramer and Bosman [30]	Cross-domain	Open Science workflow	2017
W28	Xiao [59]	Cross-domain	Reproducible Research Cycle	2021
W29	Tse et al. [53]	Medical and health sciences	Application of open science for COVID-19 vaccine/treatment development	2020
W30	Chávez Arroyo et al. [12]	Natural sciences	Open Science approach	2019
W31	Minelli et al. [34], Minelli et al. [35], Minelli et al. [36], Minelli et al. [37]	Natural sciences	Open Research Lifecycle	2017
W32	Beck et al. [8]	Natural sciences	Open science paradigm workflow	2020
W33	Reimer et al. [48]	Social sciences	Scientific workflow	2019

These terms are almost exclusive to one item, only ‘open scienc* workflow’ appears in 10 items.

The lifecycle-based family of terms stems from 14 corpus items and includes ‘open research life cycl*’, ‘open research lifecycl*’, ‘open research project lifecycl*’, ‘open scientif* process lifecycl*’, ‘reproduc* research cycl*’, ‘research cycl*’, ‘research data lifecycl*’, ‘research workflow cycl*’, and ‘scientif* life cycl*. Also in this case the terms are almost exclusive but ‘open research lifecycl*’ appearing in 5 items.

The “others” stems from 7 items opting for six different naming solutions. Four of them generically refer to an open science approach (‘application of open science’, ‘open science approach’, ‘open science paradigm’), while of the remaining three, 2 employ the nouns ‘process’ and 1 ‘phases’.

Not all the terms cite open science. Among the entries that refer to a different approach of doing research or science, 19 entries refer to ‘open science’, ‘open scientific’, or ‘open-scientific’, of which 1 also uses the qualifier ‘open scholarship’, 7 to ‘open research’, 1 to a generic ‘open’, 1 to ‘reproducible research’, 1 to a generic ‘reproducible’, and 1 to ‘collaborative scientific’. The remaining 11 entries just qualify the nouns with ‘research’ (8 entries) or ‘scientific’ (3 entries).

Such variety is an indicator of the different existing perspectives on open science and constitutes a complexity factor to be considered when comparing the effectiveness of the proposed workflows towards the achievement of open science practices (Sec. 5).

4.1.4 Scientific domains

The Frascati framework [42] was used to annotate the corpus items with respect to their primary field of science (cf. Tab. 3).

Fig. 3 depicts the resulting distribution. The majority of items (17 out of 40) does not target a specific domain while no one of the items can be considered stemming from Agriculture and veterinary science. This somehow suggests that many workflows are domain agnostic, not meant to serve the needs of a specific domain, as well as it shows that there is some interest for workflows across almost all the fields.

4.2 Workflows Features

To respond to the research question RQ5, i.e., how the proposed workflows relate to the facets characterising open science, the UNESCO Recommendation on Open Science [54] were exploited. In particular, the following six major aspects are considered.

Open infrastructures and (re)used open research products. Shared research infrastructures and open access to scientific publication, data, educational resources, software, hardware, and infrastructures are two of the open science key pillars. Consequently, one of the first aspects to consider in the workflows analysis is their use and reliance on existing services and research products.

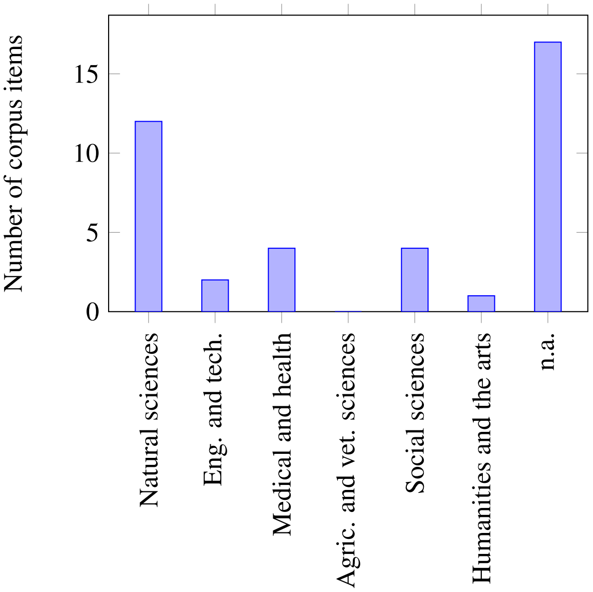


Figure 3. Corpus item field of science

Open research products created. Not every research product created during a research activity is always openly shared, if shared at all. This aspect looks at the research objects produced during the workflows that are also made openly available.

Collaboration and engagement. Open engagement of societal actors and open dialogue with other knowledge systems are among the open science foundational aspects identified by the UNESCO Recommendation. Moreover, given the importance of collaboration among researchers, we inquired if collaboration is explicitly part of the workflows and how it is pursued.

Assessment. The assessment of the research outputs is another aspect recognised by the UNESCO Recommendations as a necessity for operationalising open science and included among the open science guiding principles. Still, while the UNESCO Recommendations include it within the scope of open scientific knowledge, we regarded it as an independent parameter to highlight the processes characterising the different workflows.

Automation. The automation of the research processes is not an open science aspect per se, still it directly affects the reproducibility of the results, which is one of the guiding principles for open science identified by UNESCO. As such we considered it as a desirable aspect of an open science workflow.

Transparency. This is another concept strictly connected to open science, in turn listed among the UNESCO open science guiding principles. While transparency can have different implementation perspectives, in the context of this study the different degrees of transparency are defined in terms of which research products are openly shared and when in order to document the research processes.

While these characteristics equally apply to every workflow, the different abstraction levels that characterise the 33 workflows can result in the actual impossibility of conducting a strict comparison.

This is particularly evident in the two cases of a workflow defined at an infrastructure level. The workflows defined for diversity4bio (W20 [57]) and D4Science (W24 [4]) define in fact a range of possibilities rather than one path for the researchers to follow. Consequently the actual implementation of the workflow depends ultimately on the end-user (so that this type of workflow can be defined as user-dependent), who can choose how to use the services provided (e.g., when to actually share a research product hosted on a workspace). Still, since they provide the possibility to comply with our analysis parameters, we evaluated the characteristics exhibited by the user-dependent workflows as fully compliant with an open science approach.

4.2.1 Open infrastructures and (re)used open research products

The reuse of openly available resources is one of the advantages advocated for an open science approach.

Consequently the first aspect we evaluated is the inclusion of a 'do not reinvent the wheel' methodology, distinguishing two categories of reused resources on which the workflows are based: the *services* and the *actual research products* reused.

The first is about the use of open scientific infrastructures and it is further divided between: (*i*) physical open scientific infrastructures (e.g., wet labs), and (*ii*) virtual open scientific infrastructures (e.g., virtual research environments). Tab. 4 shows that not all the workflows mentioned the need for open science infrastructures. 8 out of 33 did not include any reference to such requirement, 25 included references to virtual open science infrastructures, including open access journals and publication platforms, bibliometrics systems, and a wide range of data services (e.g., computation, manipulation, storage, analysis), and just 3 (all included among the 25 workflows mentioning virtual open science infrastructures) can be considered as mentioning the need for physical open science infrastructures.

The second concerns the typologies of research products that are reused during the workflow, organised into four main categories: (*i*) open source software, (*ii*) open hardware, (*iii*) open research data, and (*iv*) open educational resources. The reuse of open access scientific publications could be considered as a fifth category, but its inclusion as an analysis facet did not contribute to the results in a significant manner, since the workflows can all be considered implicitly or explicitly based on the knowledge derived from existing scientific literature. Tab. 5 reports the reused resources per workflow. 18 out of 33 included references to the use of open source software and only 1 workflow considered open hardware as a required open science factor. 13 workflows out of 33 explicitly implemented the possibility of reusing open data, among which 2 also included the reuse of open educational resources (OER).

Table 4. Workflows and Open Science Infrastructures

Infrastructure	Workflows
Physical	W26 [18], W27 [28, 31, 10, 29, 30], W29 [53]
Virtual	W01 [23], W04 [17], W06 [33], W09 [46], W10 [43, 40], W11 [32], W12 [20], W13 [20], W14 [20], W15 [52], W16 [13], W18 [26], W19 [55], W20 [57], W21 [6], W22 [45], W23 [9], W24 [4], W26 [18], W27 [28, 31, 10, 29, 30], W28 [59], W29 [53], W30 [12], W32 [8], W33 [48]
n.a.	W02 [21], W03 [50, 36], W05 [16], W07 [38], W08 [24], W17 [5], W25 [19, 14], W31 [34, 35, 36, 37]

Table 5. Workflows and Reused Research Products

Reused Resource	Workflows
OS Software	W01 [23], W02 [21], W04 [17], W05 [16], W10 [43, 40], W13 [20], W15 [52], W16 [13], W17 [5], W19 [55], W21 [6], W22 [45], W23 [9], W24 [4], W26 [18], W27 [28, 31, 10, 29, 30], W28 [59], W32 [8], W33 [48]
Open Hardware	W26 [18]
Open Res. Data	W01 [23], W02 [21], W05 [16], W17 [5], W18 [26], W22 [45], W23 [9], W24 [4], W26 [18], W27 [28, 31, 10, 29, 30], W29 [53], W30 [12], W32 [8],
Open Edu. Res.	W26 [18], W27 [28, 31, 10, 29, 30],
n.a.	W03 [50, 36], W06 [33], W07 [38], W08 [24], W09 [46], W11 [32], W12 [20], W14 [20], W20 [57], W25 [19, 14], W31 [34, 35, 36, 37]

4.2.2 Open research products created

The aspects analysed for the (re)used research products generally apply also to the open research products created/released during a workflow, but, following the results of the analysis, they do not include open hardware. Tab. 6 shows the created/released research products per workflow: 28 workflows cited open research data products (including data, metadata, software parameters, tables, figures, logs, web applications, sound recordings, videos); 25 workflows cited open research products that can be broadly classified as open access scientific publications (including papers, reports, data management plans, preprints, notebooks, study designs, methods, protocols, and posters); 17 workflows cited open research software (including code, analysis scripts, and computational workflows);

6 workflows cited open educational resources (including open source tools documentation, open study material, and sharing of recorded experiments). Only 4 workflows (W19 [55], W24 [4], W25 [19, 14], and W27 [28, 31, 10, 29, 30]) mentioned the sharing of products for each one of the four categories.

Table 6. Workflows and created/released Research Products

Research Product	Workflows
OA Publication	W01 [23], W02 [21], W05 [16], W06 [33], W07 [38], W11 [32], W12 [20], W13 [20], W14 [20], W16 [13], W17 [5], W18 [26], W19 [55], W21 [6], W23 [9], W24 [4], W25 [19, 14], W26 [18], W27 [28, 31, 10, 29, 30], W28 [59], W29 [53], W30 [12], W31 [34, 35, 36, 37] W32 [8], W33 [48]
Open Res. Data	W01 [23], W03 [50, 36], W04 [17], W05 [16], W06 [33], W07 [38], W09 [46], W10 [43, 40], W11 [32], W12 [20], W13 [20], W14 [20], W16 [13], W17 [5], W18 [26], W19 [55], W20 [57], W22 [45], W23 [9], W24 [4], W25 [19, 14], W26 [18], W27 [28, 31, 10, 29, 30], W28 [59], W30 [12], W31 [34, 35, 36, 37] W32 [8], W33 [48]
Open Edu. Res.	W01 [23], W06 [33], W19 [55], W24 [4], W25 [19, 14], W27 [28, 31, 10, 29, 30]
OS Software	W04 [17], W05 [16], W12 [20], W13 [20], W14 [20], W16 [13], W17 [5], W19 [55], W22 [45], W24 [4], W25 [19, 14], W27 [28, 31, 10, 29, 30], W28 [59], W30 [12], W31 [34, 35, 36, 37] W32 [8], W33 [48]
n.a.	W08 [24], W15 [52]

4.2.3 Collaboration and engagement

The collaboration aspects taken into consideration for the analysis are divided into three groups: (*i*) the implementation solutions and their extent, (*ii*) the extent of the engagement with the society, and (*iii*) if other knowledge systems are involved.

The last two points are further expanded into four and three categories respectively, mirroring the UNESCO recommendation. Societal engagement is in fact divided into (*i*) crowdfunding, (*ii*) crowdsourcing, (*iii*) scientific volunteering, and (*iv*) citizen and participatory science. The dialogue with other knowledge systems is structured in the involvement of the three groups (*i*) indigenous people, (*ii*) marginalised scholars, and (*iii*) local communities.

The implementation of collaborative practices is explicitly cited by only 9 workflows (W01 [23], W11 [32], W14 [20], W17 [5], W20 [57], W21 [6], W24 [4], W27 [28, 31, 10, 29, 30], and W32 [8]), of which 2 (W11 and W14) just mention the possibility to collaborate and just 2 refers to a specific solution (W21 mentions the use of GitHub as a central repository, W24 mentions and the use of the social networking platform and services facilitating the actual sharing of research artefacts of the D4Science virtual research environment [4]). Among the collaboration possibilities mentioned are the use of open annotations, the creation of collaborative bibliographies, writing and coding collaboratively, and the use of immersive virtual reality.

With regards to the open engagement of societal actors (Tab. 7), only one workflow included both crowdfunding and crowdsourcing practices, 2 scientific volunteering, and 3 citizen and participatory science.

Table 7. Workflows and societal actors engagement

Research Product	Workflows
Crowdfunding	W27 [28, 31, 10, 29, 30]
Crowdsourcing	W27 [28, 31, 10, 29, 30]
Scient. volunteer.	W01 [23], W20 [57]
Citizen science	W25 [19, 14], W26 [18], W27 [28, 31, 10, 29, 30]
n.a.	W02 [21], W03 [50, 36], W04 [17], W05 [16], W06 [33], W07 [38], W08 [24], W09 [46], W10 [43, 40], W11 [32], W12 [20], W13 [20], W14 [20], W15 [52] W16 [13], W17 [5], W18 [26], W19 [55], W21 [6], W22 [45], W23 [9], W24 [4], W28 [59], W29 [53], W30 [12], W31 [34, 35, 36, 37], W32 [8], W33 [48]

Establishing an open dialogue with other knowledge systems is the least considered among the open science characteristics, with just one workflow (W27 [28, 31, 10, 29, 30]) directly involving marginalised scholars and local communities in the research process, with the latter being considered a valorisation activity rather than an actual part of the research. No references are made by any workflow to the inclusion of indigenous peoples, an aspect that is mainly linked to the geopolitical context of a research process.

4.2.4 Assessment

With regard to the assessment of the research products created during each workflow, we inquired if an open assessment phase is included and which products it affects.

An open assessment process is mentioned in the description of only 9 workflows (W06 [33], W10 [43, 40], W14 [20], W17 [5], W24 [4], W25 [19, 14], W26 [18], W27 [28, 31, 10, 29, 30], and W31 [34, 35, 36, 37]). Assessment takes very different forms, ranging from feedback in the form of comments (e.g., W14 mentioning signed comments) to a formal open peer-review process (e.g., W6, W25, W26, W27), and including open annotations (which can be considered also as a sort of weak assessment practice) (e.g., W17), transparent peer-review (e.g., W27), and non-journal organised peer review (e.g., W27).

4.2.5 Automation

The reduction of human error, the standardisation of procedures, and, ultimately, the avoidance of human bias are among the characteristics pursued by the adoption of automated workflows, that, from data acquisition to the final results, can be an invaluable instrument against the rising concern of a reproducibility crisis of scientific outputs. Following this rationale we examined if, how, and to what extent automation is part of the workflows.

12 workflows (W01 [23], W02 [21], W04 [17], W05 [16], W19 [55], W21 [6], W22 [45], W23 [9], W24 [4], W27 [28, 31, 10, 29, 30], W30 [12], and W32 [8]) considered the automation of the processes, which can be achieved partially, mainly by employing R- or Python-based notebooks, or entirely, thanks to the creation of ad-hoc solutions such as dedicated e-infrastructures and Virtual Research Environments. For instance, W19 envisages the use of an R package offering an RStudio project template allowing for automated steps, such as the synchronisation with a remote repository, the creation of a readme, the assisted creation of a preregistration and its upload on a preregistration server, data processing for generating tidy and shareable data, data analysis, and the dynamic document generation. Similarly W01 is based on the creation of a completely traceable, reproducible, and machine-actionable study package, documenting every step of the research. Once created, a machine-actionable protocol is executed using open source tools provided at the organisation level, operating on data provided through a federated database. W24 offers a data analytics platform enacting users to transform any user-defined process into an actionable process that can be executed by any other user and that automatically produces a provenance record for any execution making it fully reproducible [4].

4.2.6 Transparency

We observed that one of the most distinctive characteristics among the different workflows was their approach towards research products sharing and ultimately towards the transparency of the entire workflow. Based on what it is openly shared and when, we could in fact distinguish among different transparency degrees, as not all workflows envisage that every research product should be made openly available and not all workflows follow an ‘as soon as possible’ sharing paradigm.

With regards to the sharing timing (Tab. 8), we observed that the workflows adopted approaches that can be reduced to four models: (*i*) sharing at the end of the workflow, (*ii*) sharing part of the research products during the workflow and the rest at the end of it (mixed), (*iii*) sharing iteratively during or at the end of the related workflow phase, and (*iv*) user-dependent sharing, where it is ultimately up to the researcher to decide when to share the research products since the workflow offers different paths to follow while imposing no sharing constraint.

While it was not always possible to determine with certainty which open (or supposedly so) research products were produced or when they were shared, since not all workflows included descriptions or explicitly cited the openness of the outputs, with regard to the transparency we could nonetheless distinguish among three types of workflows (Tab. 9): (*i*) workflows with built-in transparency, (*ii*) transparency-enabled workflows, and (*iii*) opaque workflows.

Table 8. Workflows and sharing timing

Sharing timing	Workflows
At the end	W02 [21], W03 [50, 36], W04 [17], W05 [16], W08 [24], W09 [46], W12 [20], W21 [6]
Iterative	W01 [23], W13 [20], W14 [20], W16 [13], W19 [55], W22 [45], W23 [9], W25 [19, 14], W26 [18], W27 [28, 31, 10, 29, 30], W28 [59], W29 [53], W31 [34, 35, 36, 37], W32 [8], W33 [48]
Mixed	W06 [33], W10 [43, 40]
User-dependent	W20 [57], W24 [4]
n.a.	W07 [38], W11 [32], W15 [52] W17 [5], W18 [26], W30 [12],

Table 9. Workflows and transparency

Transparency	Workflows
Built-in	W01 [23], W14 [20], W16 [13], W22 [45], W23 [9], W25 [19, 14], W26 [18], W27 [28, 31, 10, 29, 30], W31 [34, 35, 36, 37], W32 [8], W33 [48]
Enabled	W19 [55], W20 [57], W24 [4]
Opaque	W02 [21], W03 [50, 36], W04 [17], W05 [16], W06 [33], W07 [38], W08 [24], W09 [46], W10 [43, 40], W11 [32], W12 [20], W13 [20], W17 [5], W18 [26], W21 [6], W28 [59], W29 [53], W30 [12]
n.a.	W15 [52]

The first group is composed of 11 workflows that implement a completely transparent open science approach at each stage, sharing every research product as soon as possible or at a later stage (in this last case the workflows must be based on standardised and openly documented processes, which, in turn, must be based on open solutions), through constraints imposed by the organisation and/or the e-infrastructure enabling the workflow, so that it is not envisaged another possibility but to follow a standardised and predefined open science process sequence. The group also include the workflows that openly share the research products or share them at a later stage, so that even if the processes are made known by the end of the workflows, they are not open to scrutiny while they are underway, but a methodology is shared before the beginning of the actual research activities and the processes are based on open solutions. This can be, for example, the case of the workflows implementing a preregistration phase (e.g., W32 [8]). While sharing the research products at a later stage, they openly share the entire methodology underlying the study at its beginning and before any other research activity but the study design, allowing for the assessment of the research processes and their results, despite at a later stage. While preregistration can be a useful transparency and reproducibility enabling driver, its effectiveness is, however, entirely dependent on the information the researchers are willing to include in it.

In the second group we included 3 workflows that can potentially be completely transparent, since the infrastructure and the service offering allow for real-time or iterative sharing of the research products at every possible moment, but it is ultimately a choice given to the end-user. This group encompasses the previously mentioned cases of the two workflows that can be defined as user-dependent (W20 [57] and W24 [4]), so that even if they can be considered fully compliant with an open science approach, the user can always opt in for a less open or even totally closed approach. In a similar manner it encompasses other workflows that, while recommending an open approach and enabling an all-in open science approach (e.g., a real-time synchronisation of the research products in an open repository like GitHub, allowing everyone to visualise, comment, and fork), also allow for a partially closed one (e.g., the use of a private repository and the sharing of the results in a following workflow stage), based on the preferences of the researchers.

Finally, the last category is composed of 18 workflows that openly share at least part of the research products created, but in doing so they do not document the processes implemented for producing them if not towards the ending stages or in outputs released with closed licences. For instance, W02 [21] and W12 [20] envisage the publication of the results in open access journals at the end of the research process, with the latter also considering the publication of data, the code used and a poster, which is based on statistics and figures that are not openly accessible.

5. Discussion

Threats to validity The most notable limitations threatening the validity of this study [3] can be summarised in the following points: (*i*) the keyword selection may not be exhaustive; (*ii*) the queries were mainly executed on titles, keywords, and abstracts; (*iii*) the number of entries of interest for this study is less than ideal for producing results that are statistically relevant; (*iv*) the study is based on the information found in the retrieved resources, which in some cases may be lacking in semantics due to their different and various types.

While the use of the term open science is widely recognised for indicating the different approach to the scientific process [39] and the term open research is attested by the literature and by its use for denoting open science related initiatives, there may be different terms employed in other contexts impacting the recall of the information retrieval process.

While there are databases that always perform a full text search, like Google Scholar, we searched mainly among titles, abstracts, and keywords (Open Research Europe does not allow searching among keywords). This choice may have limited the final number of results, but it has proved instrumental in improving the precision of the results, which already suffered from a high percentage of entries (59.9%) classified as not relevant for this study.

Suffering from this cascade effect, despite the quite large initial search space, the number of entries suitable for this study has been lower than expected.

The information carried by the different typologies of resources is quantitatively and qualitatively very different. While an image can carry the essential information about a workflow, namely the phases and their sequence, a journal article may contain descriptions and examples which can be instrumental for understanding dynamics that are not self-evident. While we tried to integrate the missing pieces of information about the workflows, it was not always possible to find additional resources about them, especially in the case of the 'other literature type' items. As a consequence the workflow comparison we developed may suffer from a discretionary interpretation, however based as strictly as possible on the information found.

Workflow shortage. The surprisingly low amount of literature on the subject and the related low number of workflows found is certainly noteworthy.

Open science is characterised by many different dimensions, entailing consequences on the research processes of great magnitude. Moreover, the research funding organisations are increasingly pushing towards the implementation of open science practices, starting with the requirements established by the European Commission for benefiting from the Horizon Europe programme. Despite that, the study of workflows implementing an open science approach has hitherto received very little attention and quite recently, given the literature yearly distribution starting from 2014.

This trend might suggest that the attention registered in the past years is the result of a process begun with the research funding organisations that is just beginning to take root in the scientific community.

Terminological diversity. The terminology regarding the conceptual representation of the workflows is indeed not standardised. Both heads and modifiers of the terms used for identifying the workflows are in fact characterised by a high variability. With regards to the heads, the terms 'workflow' and 'lifecycle' are the most used, even if not in a manner consistent with their standard acceptations. While the term 'workflow' should in fact be used to designate a series of activities [1] and 'lifecycle' for referring to the evolution of an entity [2], their actual use in the corpus items found suggests that the distinction is not always followed. As for the modifiers, while 'open' is the most used term, the designations do not always imply or refer to a change to the research process.

Open Science aspects. When analysing which open science aspects are observed in the workflows, open access to research outputs is the most represented and implemented, both in the case of reused research products and produced ones. The unequal level of attention given to the different research products and the differences in the release timelines, however, are distinctive to diverse approaches, which in turn entail different open science understandings.

Research products (re)use and sharing. Based on the explicit references made in the analysed workflows, data (including metadata), and publications (including study designs, notebooks, data management plans, preprints, journal articles, reports) are the research products whose reuse and sharing is envisaged the most in the workflows.

Open research data. Data reuse and sharing are in fact mentioned in 13 and 28 workflows respectively. With regards to the publication reuse and output, while the existing scientific literature can be considered an implicit prerequisite for the workflows, the publication output is second only to the sharing of data, with 25 mentions.

The relevance of data in the analysed workflows, even when compared to publications, can be seen as an early evidence for one of the changes entailed by an open science approach applied to the research workflow.

The supremacy of the paper as the main result of a research process, relegating the other research products to a marginal role, which is a paradigm that characterises the current evaluation dynamics, is in fact one of the aspects challenged by an open science workflow.

The importance given to the collective benefit deriving from sharing a research product as soon as possible is one of the drivers of a redistribution of merits within the research workflows, which in turn is reflected on the role and the relevance given to every research product reused or created during each phase of an open science workflow.

Open source software. While reusing existing code and sharing the produced one is a practice well represented (with 19 and 17 mentions respectively), because of the central role it takes on along with data in the fourth paradigm of science [22] and its widespread and growing adoption in scientific domains that were traditionally less accustomed to computational analysis, it would require more attention than it has to date. Moreover, there is no mention of the specifications of the environment used to compile and execute it.

Open educational resources. Among the research products, open educational resources are the least mentioned when considering the outputs of the workflows, and one of the most underrepresented open science characteristics.

The sharing of open educational resources received in turn very little attention, with only one explicit mention of the OER resource category (W25 [19, 14]) and four cases in which references were made to open study material (W19 [55]), sharing of video recording of experiments (W27 [28, 31, 10, 29, 30]), sharing of tool documentation (W01 [23]), and practice guidelines (W06 [33]) respectively.

Open hardware. While open hardware is mentioned as a reused resource, but just in one workflow, there is no mention of the sharing of open hardware as a research output of the workflows.

Open Infrastructures. The use of open infrastructure is widely mentioned as only eight workflows are generic enough not to cite any. Still there is a noteworthy difference between the references to virtual open infrastructures and to physical ones, the latter being explicitly mentioned only by two workflows. Among the cited virtual infrastructures the vast majority can be classified as repositories or as publishing platforms (including open access journals) and it is linked to the sharing of data and publications, while there is only one reference to a virtual research environment with analytics capabilities. Analytics are in fact better represented within the open software category, as the computational tasks seem to be delegated to software running in a local machine rather than to virtual infrastructures.

Research environment. As one of the goals of open science is to foster reproducibility, the environment used to conduct the research activities has to be known, if not made accessible, including the code used and the software dependencies that are required. The little attention given to the description of the research environment, as only one workflow mentioned the sharing of information on dependencies and versions, is certainly one of the criticalities that emerged from the analysis of the workflows.

Nuances of open science. In order to investigate the possible different nuances of the meaning of open science, given the complexity of the question, all of the previous analyses regarding the 20 workflows features discussed in Sec. 4.2 were aggregated into a single score to be considered jointly. Tab. 10 reports such a per workflow score calculated by counting all the features explicitly exhibited by each of the analysed exemplars (e.g., W27 exhibits the following 17 features: (*i*) use of virtual or physical infrastructures, (*ii*) use of open source software, (*iii*) use of open research data, (*iv*) use of open educational resources, (*v*) production of open access scientific publications, (*vi*) production of open source software, (*vii*) production of open research data, (*viii*) production of open educational resources, (*ix*) enabled or built-in transparency, (*x*) implementation of collaborative practices, (*xi*) crowdfunding, (*xii*) crowdsourcing, (*xiii*) citizen and participatory science, (*xiv*) dialogue with marginalised scholars, (*xv*) dialogue with local communities, (*xvi*) inclusion of open assessment practices, and (*xvii*) inclusion of automated practices)[1].

[1] The three missing features considered for the score are: use of open hardware, scientific volunteering, and dialogue with indigenous peoples

Table 10. Workflows and Open Science score

Score	Workflows
17 / 20	W27 [28, 31, 10, 29, 30]
11 / 20	W24 [4]
10 / 20	W01 [23], W26 [18]
9 / 20	W32 [8]
8 / 20	W19 [55]
7 / 20	W14 [20], W17 [5], W22 [45], W23 [9], W25 [19, 14]
6 / 20	W05 [16], W16 [13], W30 [12], W33 [48]
5 / 20	W04 [17], W06 [33], W13 [20], W20 [57], W21 [6], W28 [59], W31 [34, 35, 36, 37]
4 / 20	W02 [21], W10 [43, 40], W11 [32], W12 [20], W18 [26]
3 / 20	W29 [53]
2 / 20	W07 [38], W09 [46], W15 [52]
1 / 20	W03 [50, 36]
0 / 20	W08 [24]

When considering every single open science aspect alone, not one of them is always present in all workflows. Even the most represented aspect, open research data (including metadata), is not explicitly mentioned in 5 workflows. Therefore, it is not possible to identify a set of minimum requirements for considering a workflow in line with an open science approach.

The case of W08 [24] is of particular interest since, apparently, there are no open science aspects manifested. It is a workflow defined by the Center for Open Science that we know to be based on the Open Science Framework for enabling an open science approach. Still, it is noteworthy that the workflow phases identified can be used for defining an open science workflow as well as a fully closed one.

Such a perspective seems to suggest that open science is not a disruptive phenomenon and that the research processes do not need to be restructured or rethought in order to pursue the objectives of an open science paradigm.

Quite similar to this point of view is the one that considers open access sufficient for enabling open science, to the point that the two concepts overlap.

When considering all of the open research products created in every workflow, open accessibility is in fact the only aspect shared by the vast majority of the workflows, with the only exception of W08. Open access is certainly one of the most observed angles on open science, characterising at the very least 9 workflows (W03 [50, 36], W07 [38], W08 [24], W09 [46], W12 [20], W13 [20], W19 [55], W28 [59], W29 [53]) that do not exhibit any of the analysed open science aspects if not for the sharing of at least one open access research product.

Considering the main open science aspects declared at the beginning of Sec. 4.2, which groups similar features, it is possible to observe different combinations among them, but it is impossible to define a scale or degree of open science workflows because there is not a direct correlation among the aspects analysed. Still it is noteworthy that only W27 [28, 31, 10, 29, 30] and W24 [4] exhibit at least one desirable feature for every group.

6. Conclusion

This work surveyed the state of the art of open science workflows by analysing a corpus of 40 scientific publications, resulting in 33 unique workflows.

Six research questions were answered: (*i*) the temporal distribution characterising the publication of open science workflows occupy a period of nine years, spanning from 2014 to 2022, reflecting the recent and little attention given to the topic; (*ii*) the "means" used to publish open science workflows are varied and goes well beyond conventional publications, stressing the importance of grey literature in the matter, which is mostly considered a practical one; (*iii*) the terms used for naming open science research workflows are diverse, not standardised and not always related to the term open science; (*iv*) the scientific domains where the open science workflows originate from are almost all the existing ones with a great number of domain-agnostic workflows; (*v*) across the workflows there is a very limited sharing of common facets, open access to scientific outputs is the only common denominator; (*vi*) the different combinations of the open science aspects exhibited by the workflows delineate very different understandings of open science and, consequently, of its implementation.

Given the variety of views and solutions proposed by the analysed workflows, it is hardly possible to speak of an open science approach, rather of an open science spectrum, effectively referring to a nuance of meaning of open science. Multiple concurring concepts of open science were in fact observed, or, at the very least, many different ways the concept is perceived by researchers.

The findings effectively highlight a gap between how open science is theorised and how it is perceived and realised in practice.

Overall it is possible to argue that the analysis showed an understanding of an open science workflow characterised by the lack of structural peculiarities, as it seems that there is no particular need to define other research phases or to rearrange them in order to meet the challenges posed by open science.

While open science is being growingly advocated by research funding organisations, it is worrying to notice the little attention given to the effects of its application to the research processes, since, at least in our view, it is not just a set of principles to apply to the research workflows, rather it implies the need to rethink the processes through which knowledge is created and disseminated at a community level.

Future work will include the proposal of an open science workflow based on the identified open science aspects.

Data availability

The data that support the findings of this report are openly available on Zenodo [11].

Acknowledgments

This work received funding from the European Union's Horizon 2020 research and innovation programme under: Blue Cloud project (grant agreement No. 862409), DESIRA project (grant agreement No. 818194), EOSC-Pillar project (grant agreement No. 857650).

Author contributions

According to CRediT taxonomy, authors contributed as follows: DM performed Methodology, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization; LC and DC performed Conceptualization, Methodology, Writing - Review & Editing, Supervision, and Funding acquisition.

References

[1] ISO 10013:2021(en) Quality management systems — Guidance for documented information, .

[2] ISO/IEC 19944-1:2020(en) Cloud computing and distributed platforms Data flow, data categories and data use — Part 1: Fundamentals, .

[3] A. Ampatzoglou, S. Bibi, P. Avgeriou, M. Verbeek, and A. Chatzigeorgiou. Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. *Information and Software Technology*, 106:201–230, Feb. 2019. ISSN 0950-5849. doi: 10.1016/j.infsof.2018.10.006. URL `https://www.sciencedirect.com/science/article/pii/S0950584918302106`.

[4] M. Assante, L. Candela, D. Castelli, R. Cirillo, G. Coro, L. Frosini, L. Lelii, F. Mangiacrapa, P. Pagano, G. Panichi, and F. Sinibaldi. Enacting open science by D4Science. *Future Generation Computer Systems*, 101:555–563, Dec. 2019. ISSN 0167739X. doi: 10.1016/j.future.2019.05.063. URL `https://linkinghub.elsevier.com/retrieve/pii/S0167739X1831464X`.

[5] P. Ayris and T. Ignat. 5.1 The Empires of the Future are the Empires of the Mind' [Winston Churchill]: Defining the Role of Libraries in the Open Science Landscape, July 2017. URL `https://zenodo.org/record/3610204`.

[6] K. Bastille, S. Hardison, L. deWitt, J. Brown, J. Samhouri, S. Gaichas, S. Lucey, K. Kearney, B. Best, S. Cross, S. Large, and E. Spooner. Improving the IEA Approach Using Principles of Open Data Science. *Coastal Management*, 49(1):72–89, Jan. 2021. ISSN 0892-0753. doi: 10.1080/08920753.2021.1846155. URL `https://doi.org/10.1080/08920753.2021.1846155`. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08920753.2021.1846155.

[7] B. Baumer and D. Udwin. R markdown. *WIREs Computational Statistics*, 7(3):167–177, 2015. doi: https://doi.org/10.1002/wics.1348. URL `https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1348`.

[8] M. W. Beck, C. O'Hara, J. S. S. Lowndes, R. D. Mazor, S. Theroux, D. J. Gillett, B. Lane, and G. Gearheart. The importance of open science for biological assessment of aquatic environments. *PeerJ*, 8:e9539, July 2020. ISSN 2167-8359. doi: 10.7717/peerj.9539. URL `https://peerj.com/articles/9539`. Publisher: PeerJ Inc.

[9] M. W. Beck, Esherwoo77, G. Raulerson, S. Scolaro, and M. Burke. tbep-tech/data-management-sop: v1.0, Aug. 2021. URL `https://zenodo.org/record/5224960`.

[10] J. Bosman and B. Kramer. Of Shapes and Style: visualising innovations in scholarly communication, June 2016. URL `https://figshare.com/articles/presentation/Of_Shapes_and_Style_visualising_innovations_in_scholarly_communication/3468641/1`. Publisher: figshare.

[11] L. Candela, D. Castelli, and D. Mangione. Research workflows and open science - dataset. `https://doi.org/10.5281/zenodo.7340705`.

[12] R. A. Chávez Arroyo, J. Sanz Rodrigo, P. Gancarski, E. Cantero, F. Borbón, and P. Santos. Meso-to-microscale modelling of the Atmospheric Boundary Layer: an open-science approach, May 2019. URL `https://zenodo.org/record/3228082`.

[13] K. S. Corker. An Open Science Workflow for More Credible, Rigorous Research, Mar. 2021. URL `https://psyarxiv.com/wu6sn/`.

[14] Engineering National Academies of Sciences. *Developing a Toolkit for Fostering Open Science Practices: Proceedings of a Workshop*. Sept. 2021. ISBN 978-0-309-09361-3. doi: 10.17226/26308.

[15] B. Fecher and S. Friesike. Open Science: One Term, Five Schools of Thought. In S. Bartling and S. Friesike, editors, *Opening Science*, pages 17–47. Springer International Publishing, Cham, 2014. ISBN 978-3-319-00025-1 978-3-319-00026-8. doi: 10.1007/978-3-319-00026-8_2. URL `http://link.springer.com/10.1007/978-3-319-00026-8_2`.

[16] S. K. Firth, G. Cole, T. Kane, F. Fouchal, and T. M. Hassan. AN OPEN SCIENCE APPROACH FOR BUILDING PERFORMANCE STUDIES. Chicago, IL, Sept. 2018. URL `https://www.ashrae.org/File%20Library/Conferences/Specialty%20Conferences/`

2018%20Building%20Performance%20Analysis%20Conference%20and%20SimBuild/Papers/C097.pdf.

[17] S. K. Firth, B. Howard, and J. A. Wright. OPEN SCIENCE BUILDING STOCK MODELLING: AN EXAMPLE USING GBXML, OPENBUILDING AND ENERGYPLUS. page 8, 2018.

[18] N. J. Gownaris, K. Vermeir, M.-I. Bittner, L. Gunawardena, S. Kaur-Ghumaan, R. Lepenies, G. N. Ntsefong, and I. S. Zakari. Barriers to Full Participation in the Open Science Life Cycle among Early Career Researchers. *Data Science Journal*, 21(1):2, Jan. 2022. ISSN 1683-1470. doi: 10.5334/dsj-2022-002. URL http://datascience.codata.org/article/10.5334/dsj-2022-002/. Number: 1 Publisher: Ubiquity Press.

[19] I. Grigorov, J. Carvalho, J. Davidson, M. Donnelly, M. Elbaek, G. Franck, S. Jones, R. Melero, P. Knoth, I. Kuchma, A. Orth, N. Pontika, E. Rodrigues, and B. Schmidt. Research Lifecycle enhanced by an "Open Science by Default" Workflow, Apr. 2016. URL https://zenodo.org/record/49960.

[20] S. E. Hampton, S. S. Anderson, S. C. Bagby, C. Gries, X. Han, E. M. Hart, M. B. Jones, W. C. Lenhardt, A. MacDonald, W. K. Michener, J. Mudge, A. Pourmokhtarian, M. P. Schildhauer, K. H. Woo, and N. Zimmerman. The Tao of open science for ecology. *Ecosphere*, 6(7):art120, 2015. ISSN 2150-8925. doi: 10.1890/ES14-00402.1. URL https://onlinelibrary.wiley.com/doi/abs/10.1890/ES14-00402.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1890/ES14-00402.1.

[21] K. Harald, J. Alexander, V. Fernando, M. Carlo, M. Roberto, and N. Claudia. Application Of The Data Cube Concept For Multi Temporal Satellite Imagery. A Complete Open Science Workflow For Data Science In EO. Sept. 2017.

[22] T. Hey, S. Tansley, and K. Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Oct. 2009. ISBN 978-0-9825442-0-4. Publication Title: The Fourth Paradigm: Data-Intensive Scientific Discovery.

[23] G. Hripcsak, M. J. Schuemie, D. Madigan, P. B. Ryan, and M. A. Suchard. Drawing Reproducible Conclusions from Observational Clinical Data with OHDSI. *Yearbook of Medical Informatics*, 30(01):283–289, Aug. 2021. ISSN 0943-4747, 2364-0502. doi: 10.1055/s-0041-1726481. URL http://www.thieme-connect.de/DOI/DOI?10.1055/s-0041-1726481. Publisher: Georg Thieme Verlag KG.

[24] T. Ignat. 01 Ignat Open Science and Seachange Research 2_1, Nov. 2019. URL https://zenodo.org/record/3560726.

[25] B. A. Kitchenham. Procedures for performing systematic reviews. Technical report, Keele University, Department of Computer Science, Keele University, Kelee, UK, 07 2004. URL http://www.it.hiof.no/~haraldh/misc/2016-08-22-smat/Kitchenham-Systematic-Review-2004.pdf.

[26] J. Klenk, P. R. O. Payne, R. Shrestha, and M. Edmunds. Open Science and the Future of Data Analytics. In M. Edmunds, C. Hass, and E. Holve, editors, *Consumer Informatics and Digital Health: Solutions for Health and Health Care*, pages 337–357. Springer International Publishing, Cham, 2019. ISBN 978-3-319-96906-0. doi: 10.1007/978-3-319-96906-0_18. URL https://doi.org/10.1007/978-3-319-96906-0_18.

[27] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay, et al. *Jupyter Notebooks-a publishing format for reproducible computational workflows.*, volume 2016. 2016.

[28] B. Kramer and J. Bosman. Open Science practices (editable powerpoint slides), Feb. 2017. URL https://figshare.com/articles/presentation/Open_Science_practices_editable_powerpoint_slides_/4627999/1. Publisher: figshare.

[29] B. Kramer and J. Bosman. Wheel of Open Science practices (image), Feb. 2017. URL https://figshare.com/articles/figure/Wheel_of_Open_Science_practices_image_/4628014/2. Publisher: figshare.

[30] B. Kramer and J. Bosman. Wheel of Open Science practices (editable powerpoint), Feb. 2017. URL https://figshare.com/articles/presentation/Wheel_of_Open_Science_practices_editable_powerpoint_/4628023/1. Publisher: figshare.

[31] I. Labastida i Juan. The role of an academic library assessing research in the open science age. Jan. 2017. URL http://diposit.ub.edu/dspace/handle/2445/106182. Accepted: 2017-01-30T09:40:30Z.

[32] L. Lahti, M. Tolonen, K. Lindén, and M. Paavolainen. 12.2 Open and Reproducible Analytical Workflows in the Humanities: The Case of Finnish Bibliographic Metadata, June 2015. URL https://zenodo.org/record/3603218.

[33] C. Linehan, T. Araten-Bergam, J. Baumbusch, J. Beadle-Brown, C. Bigby, G. Birkbeck, V. Bradley, M. Brown, F. Bredewold, M. Chirwa, J. Cui, M. G. Gimenez, T. Gomiero, S. Kanova, T. Kroll, M. MacLachlan, B. Mirfin-Veitch, J. Narayan, F. Nearchou, A. Nolan, M.-A. O'Donovan, F. H. Santos, J. Siska, T. Stainton, M. Tideman, and J. Tossebro. COVID-19 IDD: A global survey exploring family members' and paid staff's perceptions of the impact of COVID-19 on individuals with intellectual and developmental disabilities and their caregivers. Technical Report 3:39, HRB Open Research, Dec. 2020. URL https://hrbopenresearch.org/articles/3-39. Type: article.

[34] A. Minelli, A. Sarretta, A. Oggioni, A. Pugnetti, M. Bastianini, F. B. Aubry, T. Scovacricchi, E. Camatti, and G. Socal. Improving marine ecological data lifecycle through Open Science Principles. Apr. 2017. doi: 10.6084/m9.figshare.4822942.v2. URL https://figshare.com/articles/poster/poster_AS_AM_3_pdf/4822942/2. Publisher: figshare.

[35] A. Minelli, A. Oggioni, A. Pugnetti, A. Sarretta, M. Bastianini, C. Bergami, F. B. Aubry, E. Camatti, T. Scovacricchi, and G. Socal. The project EcoNAOS: vision and practice towards an open approach in the Northern Adriatic Sea ecological observatory. *Research Ideas and Outcomes*, 4:e24224, Feb. 2018. ISSN 2367-7163. doi: 10.3897/rio.4.e24224. URL https://riojournal.com/article/24224/. Publisher: Pensoft Publishers.

[36] A. Minelli, A. Oggioni, A. Pugnetti, A. Sarretta, M. Bastianini, C. Bergami, F. Bernardi Aubry, E. Camatti, T. Scovacricchi, and G. Socal. Figure 3 from: Minelli A, Oggioni A, Pugnetti A, Sarretta A, Bastianini M, Bergami C, Bernardi Aubry F, Camatti E, Scovacricchi T, Socal G (2018) The project EcoNAOS: vision and practice towards an open approach in the Northern Adriatic Sea ecological observatory. Research Ideas and Outcomes 4: e24224. https://doi.org/10.3897/rio.4.e24224, Apr. 2018. URL https://zenodo.org/record/1223705.

[37] A. Minelli, A. Sarretta, A. Oggioni, C. Bergami, M. Bastianini, F. Bernardi Aubry, E. Camatti, and A. Pugnetti. Opening Marine Long-Term Ecological Science: Lesson Learned From the LTER-Italy Site Northern Adriatic Sea. *Frontiers in Marine Science*, 8, 2021. ISSN 2296-7745. URL https://www.frontiersin.org/articles/10.3389/fmars.2021.659522.

[38] E. Morissette, L. Harper, I. Peters, F. Tayler, and S. Haustein. Data Management Plan Template: Open Science Workflows. Apr. 2021. doi: 10.5281/zenodo.4701021. URL https://zenodo.org/record/4701021. Publisher: Zenodo.

[39] G. Mosconi, Q. Li, D. Randall, H. Karasti, P. Tolmie, J. Barutzky, M. Korn, and V. Pipek. Three Gaps in Opening Science. *Computer Supported Cooperative Work (CSCW)*, 28(3):749–789, June 2019. ISSN 1573-7551. doi: 10.1007/s10606-019-09354-z. URL https://doi.org/10.1007/s10606-019-09354-z.

[40] S. Muftic. Automated preservation workflows: How not to think about preservation every day by thinking about it a lot on the first day, Nov. 2021. URL https://figshare.com/articles/presentation/Automated_preservation_workflows_How_not_to_think_about_preservation_every_day_by_thinking_about_it_a_lot_on_the_first_day/16929532/1. Publisher: University of Cape Town.

[41] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. L. Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. Promoting an open research culture. *Science*, 348(6242):1422–1425, June 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aab2374. URL https://www.science.org/doi/10.1126/science.aab2374.

[42] OECD. *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*. The Measurement of Scientific, Technological and Innovation Activities. OECD, Oct. 2015. ISBN 978-92-64-23880-0 978-92-64-23901-2 978-92-64-26208-9. doi: 10.1787/9789264239012-en. URL https://www.oecd-ilibrary.org/science-and-technology/frascati-manual-2015_9789264239012-en.

[43] Open Science and Research Initiative. The Open Science and Research Handbook. Technical report, Dec. 2014. URL https://www.fosteropenscience.eu/sites/default/files/pdf/3986.pdf.

[44] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson. Systematic Mapping Studies in Software Engineering. June 2008. doi: 10.14236/ewic/EASE2008.8. URL https://scienceopen.com/document?vid=6d552894-2cc3-4e2b-a483-41fa48a37ef8.

[45] D. Pieper. INTACT – Collecting data on fee-based Open Access publishing. *Septentrio Conference Series*, (5), Nov. 2015. ISSN 2387-3086. doi: 10.7557/5.3674. URL https://septentrio.uit.no/

index.php/SCS/article/view/3674. Number: 5.

[46] M. Puren and C. Riondet. Research data management, a chance for Open Science. Methods and tutorials to create a Data Management Plan (DMP). Oct. 2016. URL https://hal.inria.fr/hal-01416978.

[47] R. Ramachandran, K. Bugbee, and K. Murphy. From Open Data to Open Science. *Earth and Space Science*, 8(5), May 2021. ISSN 2333-5084, 2333-5084. doi: 10.1029/2020EA001562. URL https://onlinelibrary.wiley.com/doi/10.1029/2020EA001562.

[48] C. B. Reimer, Z. Chen, C. Bundt, C. Eben, R. E. London, and S. Vardanian. Open Up – the Mission Statement of the Control of Impulsive Action (Ctrl-ImpAct) Lab on Open Science. *Psychologica Belgica*, 59(1):321–337, Aug. 2019. ISSN 2054-670X. doi: 10.5334/pb.494. URL http://www.psychologicabelgica.com/article/10.5334/pb.494/. Number: 1 Publisher: Ubiquity Press.

[49] J. Rüegg, C. Gries, B. Bond-Lamberty, G. J. Bowen, B. S. Felzer, N. E. McIntyre, P. A. Soranno, K. L. Vanderbilt, and K. C. Weathers. Completing the data life cycle: using information management in macrosystems ecology research. *Frontiers in Ecology and the Environment*, 12 (1):24–30, Feb. 2014. ISSN 1540-9295, 1540-9309. doi: 10.1890/120375. URL https://onlinelibrary.wiley.com/doi/abs/10.1890/120375.

[50] A. Sarretta. Research Data Life Cycle, Jan. 2018. URL https://zenodo.org/record/1149049. Publisher: Zenodo.

[51] C. Tenopir, S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame. Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, 6(6):e21101, June 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0021101. URL https://dx.plos.org/10.1371/journal.pone.0021101.

[52] S. Teplitzky. Hanging out your open science shingle: launching an open science program by supporting a new graduate cohort, Oct. 2019. URL https://zenodo.org/record/3510306.

[53] E. G. Tse, D. M. Klug, and M. H. Todd. Open science approaches to COVID-19. Technical Report 9:1043, F1000Research, Aug. 2020. URL https://f1000research.com/articles/9-1043. Type: article.

[54] UNESCO. UNESCO Recommendation on Open Science. Technical report, UNESCO, Paris, 2021. URL https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en.

[55] C. J. Van Lissa, A. M. Brandmaier, L. Brinkman, A.-L. Lamprecht, A. Peikert, M. E. Struiksma, and B. M. I. Vreede. WORCS: A workflow for open reproducible code in science. *Data Science*, 4(1):29–49, Jan. 2021. ISSN 2451-8484. doi: 10.3233/DS-210031. URL https://content.iospress.com/articles/data-science/ds210031. Publisher: IOS Press.

[56] R. Vicente-Saez and C. Martinez-Fuentes. Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, 88:428–436, July 2018. ISSN 01482963. doi: 10.1016/j.jbusres.2017.12.043. URL https://linkinghub.elsevier.com/retrieve/pii/S0148296317305441.

[57] E. Wandl-Vogt, D. Ostojic, B. Piringer, H. Rainer, and K. Zsaytseva. Designing Collaborative Ecosystems and community organization: Introducing the multidisciplinary portal on “Biodiversity and Linguistic Diversity: A Collaborative Knowledge Discovery Environment”. 2017. URL https://dh2017.adho.org/abstracts/531/531.pdf.

[58] H. Wickham and G. Grolemund. *R for data science: import, tidy, transform, visualize, and model data.* ” O’Reilly Media, Inc.”, 2016.

[59] J. Xiao. Panel Discussion 1 | Investing in open infrastructure and service: The benefits and challenges at University of Hong Kong, Oct. 2021.

[60] Y. Xie. knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, and R. D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. URL http://www.crcpress.com/product/isbn/9781466561595. ISBN 978-1466561595.

Appendix I. Research Workflows Overview

This appendix provides an overview of the 33 unique workflows this report is based on. Because of the heterogeneity of the sources and the resulting significant differences in the accompanying documentation, if present, the description level may vary considerably.

W01 by Hripcsak et al. [23]

Hripcsak et al. [23] describe an infrastructure-enabled research flow, adopted by “The Observational Health Data Sciences and Informatics (OHDSI)” initiative, aimed at increasing reproducibility of the generated clinical evidence through Open Science (Fig. 4).

It is a community-oriented research workflow, based on a federated data model and a common set of open source tools and workflows used for analysing data, which guarantee the transparency of the processes within the boundaries set by the patient-level data confidentiality. The research flow begins at a community level with a data preprocessing phase (‘open community data standards’), consisting in the conversion of each organisation’s data into the OHDSI Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). The researchers publish the study design and then proceed with the analyses, which are conducted using the open source tools and workflows provided by the community. The analyses run locally, on the organisations’ infrastructures, so that their patient data can remain private, but the results are made available at the community level in the form of statistical data. Following a phase of collaborative interpretation of the results, the findings and the software parameters are publicly shared at data.OHDSI.org, while the code is shared on GitHub.

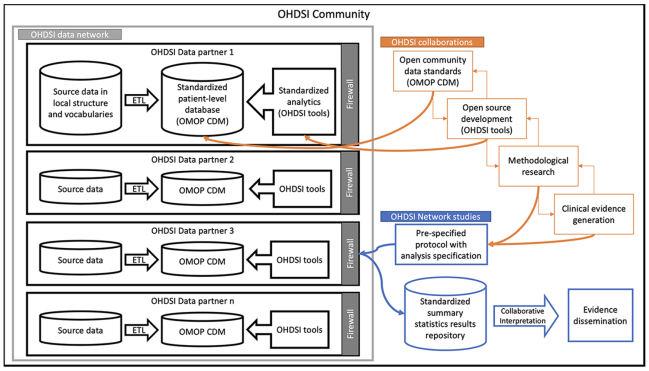

OHDSI Community
OHDSI data network
OHDSI Data partner 1
Source data in local structure and vocabularies
ETL
Standardized patient-level database (OMOP CDM)
Standardized analytics (OHDSI tools)
Firewall
OHDSI Data partner 2
Source data
ETL
OMOP CDM
OHDSI tools
Firewall
OHDSI Data partner 3
Source data
ETL
OMOP CDM
OHDSI tools
Firewall
OHDSI Data partner n
Source data
ETL
OMOP CDM
OHDSI tools
Firewall
OHDSI collaborations
Open community data standards (OMOP CDM)
Open source development (OHDSI tools)
Methodological research
Clinical evidence generation
OHDSI Network studies
Pre-specified protocol with analysis specification
Standardized summary statistics results repository
Collaborative Interpretation
Evidence dissemination

Figure 4. OHDSI Research Flow [23]

W02 by Harald et al. [21]

Harald et al. [21] present an Open Science workflow applied within the scope of the European Space Agency project SEOM - S14SCI Land “SinCohMap: Exploitation of Sentinel-1 InSAR Coherence for Land-Cover and Vegetation Mapping”. It is based on the datacube concept, as implemented by the free and open source software Rasdaman, the free and open data provided by the satellite Sentinel-2, and the use of Jupyter Notebooks for data analysis. The main feature of this workflow is that data preprocessing is centralised and carried out by a data provider. Data is then made available to researchers, at least at a project consortium level, fostering comparable EO research and ultimately reproducibility.

W03 by Sarretta [50] and Minelli et al. [36]

This workflow specifies the ‘describe’ and ‘preserve/share’ phases of the circular lifecycle defined by Rüegg et al. [49] by focusing on metadata (Fig. 5).

The described research data lifecycle is composed of three groups of phases. The first describes the lifecycle of a traditional project and the creation of a dataset. It consists of four phases, namely ‘plan’, ‘collect’, ‘quality assurance/quality control’, and ‘analyze’. The two phases ‘describe’ and ‘preserve/share’ in the second group are dedicated to describing data through metadata and preserving both data and metadata for long-term availability and re-use. The last group is composed of the three phases needed for re-using already existing data in other projects, namely ‘discover’, ‘integrate’, and ‘analyze’.

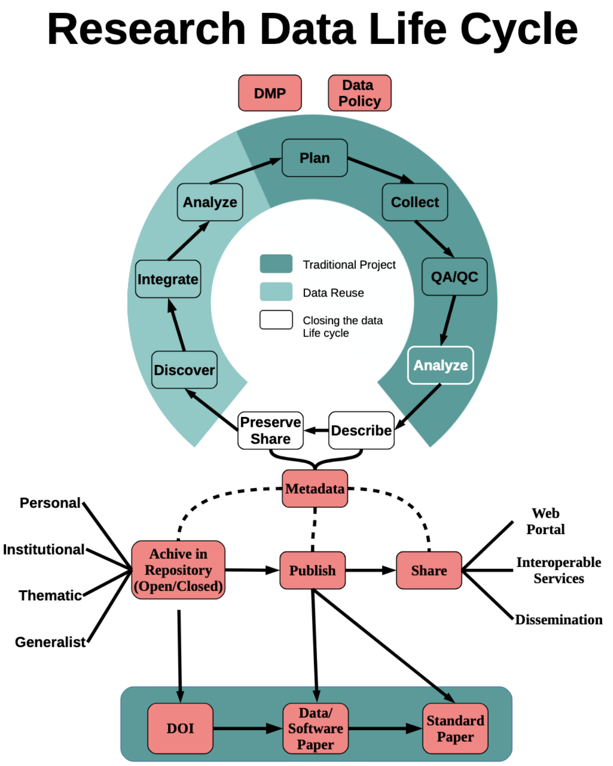


Figure 5. Research Data Lifecycle [50]

W04 by Firth et al. [17]

Firth et al. [17] describe an Open Science workflow for building stock modelling studies, consisting in predicting the energy and environmental performance of large groups of buildings through computer models. The workflow, based on a typical example of stock model type, consists of four stages, namely ‘Original datasets sourced’, ‘Original data converted to model inputs’, ‘Simulations run and results generated’, and ‘Analysis of results and findings generated’, and it is accompanied by six Open Science recommendations. In order to follow an Open Science approach within the first stage, the databases on which stock models rely should be openly available. In the second stage the buildings should be described using a model- and software-neutral data format in order to avoid the loss of important parameters and other information due to the conversion to one specific input file, and the conversion should be pursued through documented and automated means to ensure reproducibility. In the third stage the simulation should be based on open source software, either existing or created ad hoc, and the simulation results should be published in an interoperable data format. In the final stage the analysis should be conducted using well-documented, understandable and reproducible computer code, avoiding any manual intervention. Following these recommendations, in the presented workflow instance the initial data are encoded in the gbXML standard, documented Jupyter Notebooks are used for the analysis, and all data, methods, and notebooks are shared on GitHub.

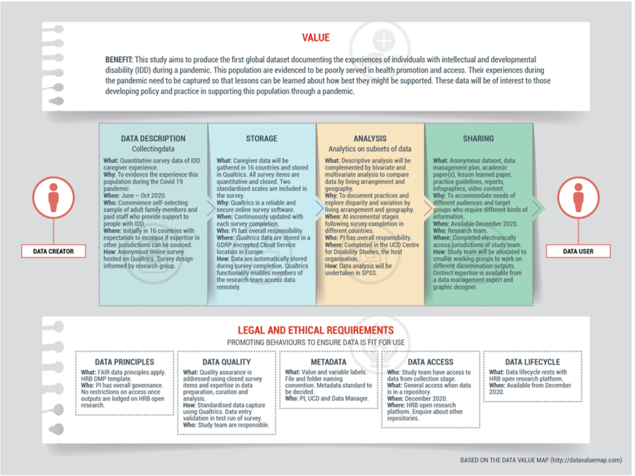
W05 by Firth et al. [16]

Firth et al. [16] describe an Open Science workflow in three phases for building performance studies. For each phase it specifies which subphases it includes, the Open Science approach that should be followed, and where the resulting research products should be published. The first phase, 'Dataset', includes the description of the building and the objects within the building and the sensor measurements. The second phase 'Analysis / Simulation' includes modelling, simulation, statistical analysis, optimization, and parametric analysis. The last phase, 'Published results', includes creating tables and generating figures and plots. In the dataset phase six points are identified for enabling an Open Science approach: (*i*) the dataset collection methods should be documented; (*ii*) the dataset should not include manually added, ambiguous, or poorly defined entries; (*iii*) the dataset should be represented in a formal language (e.g., XML, JSON); (*iv*) parsers for the representation language should exist; (*v*) the schema should be richly described (EnergyPlus idd, gbXML); and (*vi*) the dataset should be encoded in a machine-readable format. Five Open Science points are identified for the second phase: (*i*) analysis and simulations should be conducted using automated scripts to ensure reproducibility; (*ii*) there should be no manual intervention; (*iii*) the script should follow a formal and well described language (e.g., Python, R); (*iv*) the scripts should be easy to read; and (*v*) each step should be documented (e.g., using Jupyter Notebooks). Finally, six Open Science points are identified in the 'Published results' phase, three about the approach and three about the added benefit of adopting Open Science: (*i*) the results should be open access; (*ii*) the dataset and the code should be published contextually to the paper; (*iii*) all stages of the workflow should be openly available; (*iv*) the study is fully reproducible; (*v*) the increase in value of the study due to the sharing of the dataset and the methodology; and (*vi*) the benefit to the work visibility and, consequently, the increased citations. The results of each phase should be published in open access repositories, journals, and proceedings. In particular: the dataset should be published in open access data repositories and as supplement in open access journals; the results of the analysis / simulation phase and the code used should be published in open access data repositories, as supplement in open access journals, and in code repositories (e.g., GitHub); and the paper should be published in open access journals or open access conference proceedings.

W06 by Linehan et al. [33]

Linehan et al. [33] present an open research lifecycle defined within the scope of intellectual and developmental disability studies (Fig. 6). Aiming to create the first dataset on the experiences of individuals with intellectual and developmental disability during the Covid19 pandemic, the paper describes an open research lifecycle derived from the data value map[2] and based on the use of the Health Research Board (HRB) open research platform [3].

It consists of four phases, 'Data description', 'Storage', 'Analysis', and 'Sharing', that illustrate the process of bringing the data from the researchers to the public. The dataset is the result of an anonymous survey, realised using the Qualtrics Core XMTM[4] platform for data collection and storage. The IBM SPSS Version 26 statistical software is used for data analysis and the HRB open research platform is used for sharing the study outputs in the last lifecycle phase.



VALUE

BENEFIT: This study aims to produce the first global dataset documenting the experiences of individuals with intellectual and developmental disability (IDD) during a pandemic. This population are evidenced to be poorly served in health promotion and access. Their experiences during the pandemic need to be captured so that lessons can be learned about how best they might be supported. These data will be of interest to those developing policy and practice in supporting this population through a pandemic.

DATA DESCRIPTION
Collectingdata

What: Quantitative survey data of IDD caregiver experience.
Why: To evidence the experience this population during the Covid 19 pandemic
When: June – Oct 2020.
Who: Convenience self-selecting sample of adult family members and paid staff who provide support to people with IDD.
Where: Initially in 16 countries with expectation to increase if expertise in other jurisdictions can be sourced.
How: Anonymous online survey hosted on Qualtrics. Survey design informed by research group.

STORAGE

What: Caregiver data will be gathered in 16 countries and stored in Qualtrics. All survey items are quantitative and closed. Two standardised scales are included in the survey.
Why: Qualtrics is a reliable and secure online survey software.
When: Continuously updated with each survey completion.
Who: PI has overall responsibility.
Where: Qualtrics data are stored in a GDRP encrypted Cloud Service location in Europe.
How: Data are automatically stored during survey completion. Qualtrics functionality enables members of the research team access data remotely.

ANALYSIS
Analytics on subsets of data

What: Descriptive analysis will be complemented by bivariate and multivariate analysis to compare data by living arrangement and geography.
Why: To document practices and explore disparity and variation by living arrangement and geography.
When: At incremental stages following survey completion in different countries.
Who: PI has overall responsibility.
Where: Completed in the UCD Centre for Disability Studies, the host organisation.
How: Data analysis will be undertaken in SPSS.

SHARING

What: Anonymous dataset, data management plan, academic paper(s), lesson learned paper, practice guidelines, reports, infographics, video content.
Why: To accommodate needs of different audiences and target groups who require different kinds of information.
When: Available December 2020.
Who: Research team.
Where: Completed electronically across jurisdictions of study team.
How: Study team will be allocated to smaller working groups to work on different dissemination outputs. Distinct expertise is available from a data management expert and graphic designer.

LEGAL AND ETHICAL REQUIREMENTS
PROMOTING BEHAVIOURS TO ENSURE DATA IS FIT FOR USE

DATA PRINCIPLES
What: FAIR data principles apply. HRB DMP template.
Who: PI has overall governance. No restrictions on access once outputs are lodged on HRB open research.

DATA QUALITY
What: Quality assurance is addressed using closed survey items and expertise in data preparation, curation and analysis.
How: Standardised data capture using Qualtrics. Data entry validation in test run of survey.
Who: Study team are responsible.

METADATA
What: Value and variable labels. File and folder naming convention. Metadata standard to be decided.
Who: PI, UCD and Data Manager.

DATA ACCESS
Who: Study team have access to data from collection stage.
What: General access when data is in a repository.
When: December 2020.
Where: HRB open research platform. Enquire about other repositories.

DATA LIFECYCLE
What: Data lifecycle rests with HRB open research platform.
When: Available from December 2020.

BASED ON THE DATA VALUE MAP (http://datavaluemap.com)

Figure 6. Data management plan for coronavirus disease 2019 (COVID-19) intellectual and developmental disabilities (IDD) research project [33]

W07 by Morissette et al. [38]

Morissette et al. [38] present a data management plan template implementing an open science/open scholarship workflow. It defines six steps, each with multiple requirements, for ensuring data re-use: 'Responsibilities and Resources', 'Data collection', 'Documentation and metadata', 'Storage and Backup', 'Sharing, Reuse and Preservation', and 'Ethics and Legal Compliance'. The first phase deals with the indication of the individual or an organisation responsible for the data management during and beyond a project and the resources (human, hardware, software) needed. The second prescribes the use of open or industry standard formats for encoding data, the adoption of a file naming convention, and the creation of an onboarding document for standardising the workflow among the project participants. The third encourages the adoption of a readme file for documenting each

[2] Data value map `http://datavaluemap.com`
[3] HRB Open Research `https://hrbopenresearch.org/`
[4] Qualtrics Services`https://www.qualtrics.com/uk/`

dataset, specifying the information that should be documented (“information about the study, data-level descriptions, and any other contextual information required to make the data usable by other researchers” and other information including “research methodology used, variable definitions, vocabularies, classification systems, units of measurement, assumptions made, format and file type of the data, a description of the data capture and collection methods, explanation of data coding and analysis performed (including syntax files)”) and who will be in charge of the task, prescribes the adoption of a metadata standard and the identification of the secondary data sources. The fourth stage covers the storage needs, backup schedule and the use of data transport solutions. The fifth suggests the use of data repositories, in particular for long term preservation, and the identification of the data typologies to share (raw, processed, analysed, final) in light of possible legal restrictions, recommends the use of open licences for enabling data re-use, and propose the usage of different dissemination solutions (e.g., social media, forums). Finally, The last stage prescribes to identify the ethical and legal requirements to comply with in order to share the data, including privacy concerns.

W08 by Ignat [24]

Ignat [24] shows a research lifecycle defined by the Center for Open Science and implemented in the Open Science Framework (Fig. 7). It is structured in ten sequential phases: (*i*) Search and Discover; (*ii*) Develop Idea; (*iii*) Design Study; (*iv*) Acquire Materials; (*v*) Collect Data; (*vi*) Store Data; (*vii*) Analyze Data; (*viii*) Interpret findings; (*ix*) Write Report; and (*x*) Publish Report.

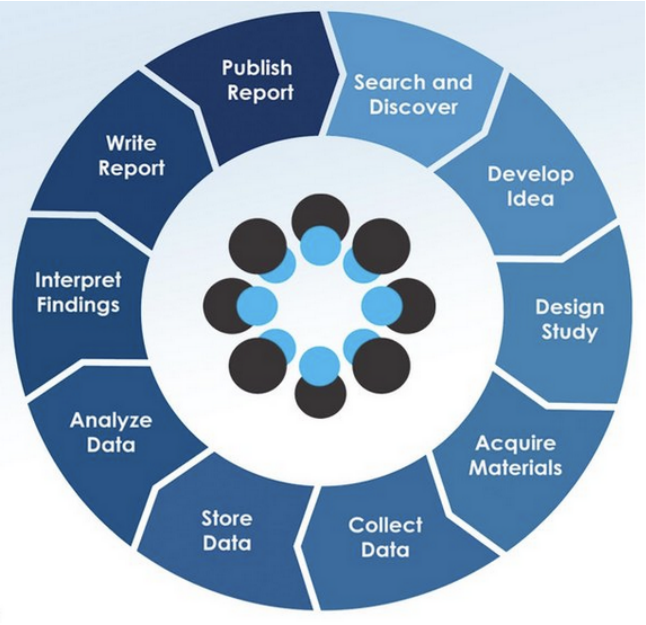


Figure 7. OSF-based workflow [24]

W09 by Puren and Riondet [46]

Puren and Riondet [46] show a circular research data lifecycle in the context of proposing a new model of openness for research data to implement in a data management plan. The research data lifecycle, created by the University of Sydney Library, consists of five phases, namely ‘Plan & Fund’, ‘Collect & Analyse’, ‘Preserve & Store’, ‘Publish & Share’, and ‘Discover & Reuse’. As part of the broader data management processes, data management plans are presented as instrumental in ensuring the reusability of the data, in particular when they are written conforming to the FAIR principles. Responsibilities for each role in the data lifecycle should be declared. Reused data should be sourced and data should be described (including their context, purpose, type, provenance, and formats) as should be the technical details of the processes involving them. File naming should avoid any ambiguous or inscrutable formulations, as well as special characters. Data should be stored in a single place, using an open format and avoiding multiple versions. Formats and standards followed should be in line with the FAIR principles and ad hoc practices are to be documented. Data to retain, share and/or preserve should be selected also taking into consideration the legal aspects and data repositories should be used for preservation and sharing purposes. Data access policies should be described and ethical and legal requirements are to be identified and declared.

W10 by Open Science and Research Initiative [43], Muftic [40]

Open Science and Research Initiative [43] propose an open approach to the research process, identifying a circular workflow (Fig. 8) composed of seven stages (‘Hypothesis’, ‘Data collection’, ‘Processing’, ‘Storing data and results’, ‘Long-term preservation’, ‘Publication and distribution’, and ‘Reuse’) and the related actions to be carried out for achieving openness. Beginning from the ‘Hypothesis’ phase, financiers’ requirements should be considered. In the ‘Data collection’ phase usage rights should be clarified and credits should be given in the case of data reuse. In the ‘Processing’ phase open software and open interfaces should be used. The use of a service infrastructure, the attachment of a PID and descriptive metadata to the results, and the publication of metadata with an open licence are recommended in the ‘Storing data and results’ stage. For the ‘Long-term preservation’ stage, services that safeguard the preservation and integrity of materials should be used and the produced metadata should follow a standard. In the ‘Publication and distribution’ stage the metadata should be published with an open licence, the evaluation should be open, the publication and the related data and methods should be linked together, and institutional repositories should be used. Lastly, in the ‘Reuse’ phase the citations made should be clear and the accumulation of credits should be ensured.

W11 by Lahti et al. [32]

Lahti et al. [32] present an open and reproducible Workflow in the Humanities for Research & Education and, in particular,

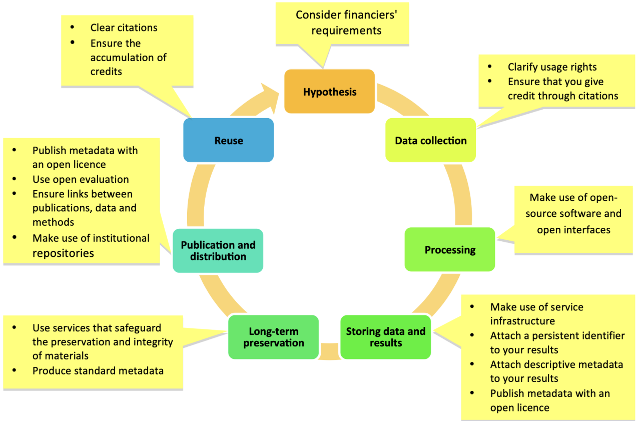

Consider financiers' requirements
Hypothesis
Clear citations
Ensure the accumulation of credits
Clarify usage rights
Ensure that you give credit through citations
Reuse
Data collection
Publish metadata with an open licence
Use open evaluation
Ensure links between publications, data and methods
Make use of institutional repositories
Make use of open-source software and open interfaces
Publication and distribution
Processing
Use services that safeguard the preservation and integrity of materials
Produce standard metadata
Long-term preservation
Storing data and results
Make use of service infrastructure
Attach a persistent identifier to your results
Attach descriptive metadata to your results
Publish metadata with an open licence

Figure 8. Promoting openness at different stages of the research process by the Open Science and Research Initiative [43]

in the case of the English Short-Title Catalogue metadata used for providing transparent quantitative analysis of knowledge production (Fig. 9).

The workflow is composed of four phases, namely ‘Preprocessing’, ‘Analysis’, ‘Visualization’, and ‘Document generation’, each one linked to the research products produced as output. Starting from the raw data, tidy data, statistical summaries, tables and figures, and the online report (to be published on Github) respectively, are in fact identified. The whole workflow is based on collaboration and transparency in the matters of data, methods, and reporting.

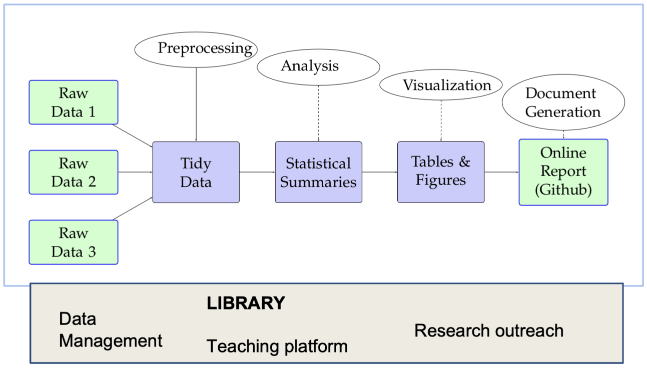


Figure 9. Reproducible Workflows in humanities for Research & Education [32]

W12, W13, and W14 by Hampton et al. [20]

Hampton et al. [20] present three examples of possible open science workflows, contextualised in the field of ecology, characterised by a different degree of openness (Fig. 10).

The first one identifies five sequential stages – ‘Designs experiment’, ‘Fieldwork’, ‘Data capture’, ‘Analysis (statistics and figures) with R’, and ‘Writes paper in Word’ – that are not open to researchers that are not directly involved in the research, but that still involve external influences to a certain extent and whose outputs are openly shared towards the end of the workflow. In the ‘Designs experiment’ phase the external influences are identified in the reading of blog posts and in the questions asked on social media, process, this last one, that creates a reciprocal relation between the researchers working on the project and the external ones. The research outputs are shared during the ‘Analysis (statistics and figures) with R’ phase, in the form of a presentation of a poster and its sharing on figshare, and in the ‘Writes paper in Word’ phase, since the paper is then submitted to Ecosphere, leading its publication, data are submitted to Ecological Archives, and the code is published as a supplement. Finally the publication link is shared in a blog post, which can be in turn an external influence for other researchers in the ‘Designs experiment’ phase.

The second example is a workflow where only the first three phases are closed: ‘Designs experiment’, ‘Writes grant proposal’, and ‘Lab work’. Like in the previous example, the ‘Design experiment’ phase is characterised by external influences, in this case the reading of a paper. The outputs of the ‘Lab work’ phase, environmental data and genetic sequence data, are then shared on the Knowledge Network for Biocomplexity data repository and the National Center for Biotechnology Information respectively. The ‘Analysis and writing with Rmarkdown, in public repository on GitHub’ follows the ‘Lab work’ phase. In addition to being open to the external researchers, it establishes a reciprocal relation between them and those directly involved in the project by the means of a talk, whose slides are shared on figshare. While the preprint is uploaded on bioRXiv, the paper, citing the related data and code, is submitted to PLoS Biology and published, becoming in turn a possible external influence in another workflow.

The last example presents a fully open workflow. It starts from the ‘Proposes idea online in open lab notebook’, which is influenced by the existing open-access publications, and continues with the phase ‘Drafts proposal, designs experiment online’, benefiting from the reader’s comments. The next phases are ‘Publishes data as it is collected’, ‘Publishes R code as package on GitHub’, ‘Publishes preprint, submits to journal, cites code and data’, which is subject to open peer-review, and, lastly, ‘Publishes open-access paper’.

W15 by Teplitzky [52]

Teplitzky [52] presents an open science workflow as part of a pilot project of the University of California’s library for fostering open science practices and reproducibility within the Earth Sciences (Fig. 11).

The workflow consists of six cycles, ‘Discovery’, ‘Data collection’, ‘Analysis’, ‘Writing’, ‘Publication,’ and ‘Outreach & impact’, in which reproducibility is pursued by adopting open source tools (e.g., Overleaf in the ‘Writing’ phase and Jupyter Notebooks throughout ‘Analysis’).

W16 by Corker [13]

Corker [13] proposes an Open Science Workflow in the con-

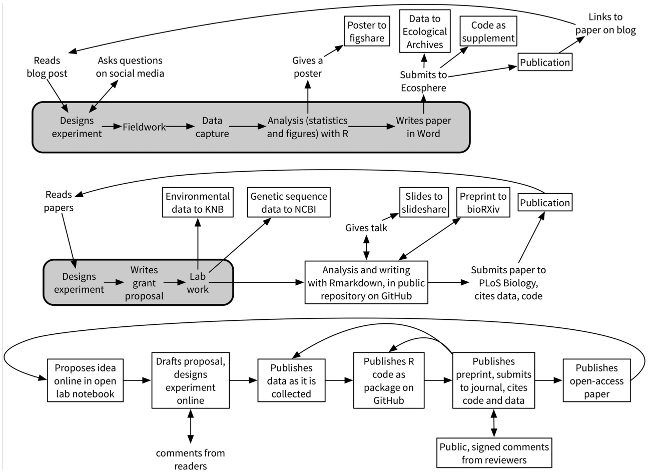
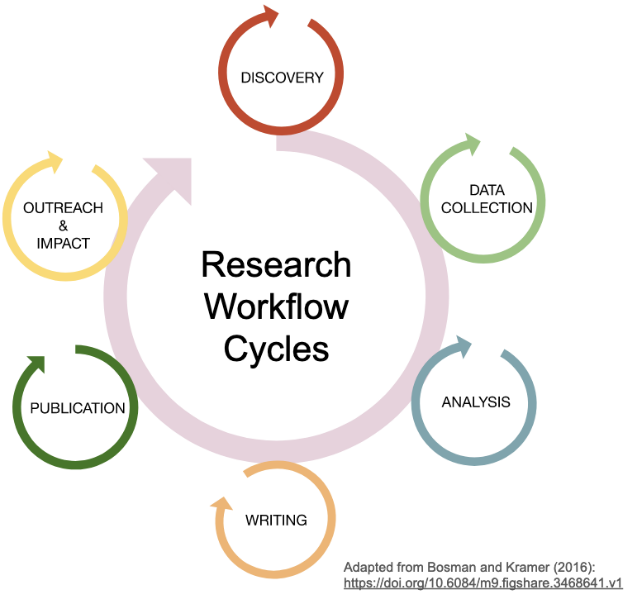

Reads blog post
Asks questions on social media
Gives a poster
Poster to figshare
Data to Ecological Archives
Code as supplement
Links to paper on blog
Publication
Submits to Ecosphere
Designs experiment
Fieldwork
Data capture
Analysis (statistics and figures) with R
Writes paper in Word
Reads papers
Environmental data to KNB
Genetic sequence data to NCBI
Slides to slideshare
Preprint to bioRXiv
Publication
Gives talk
Designs experiment
Writes grant proposal
Lab work
Analysis and writing with Rmarkdown, in public repository on GitHub
Submits paper to PLoS Biology, cites data, code
Proposes idea online in open lab notebook
Drafts proposal, designs experiment online
Publishes data as it is collected
Publishes R code as package on GitHub
Publishes preprint, submits to journal, cites code and data
Publishes open-access paper
comments from readers
Public, signed comments from reviewers

Figure 10. Three examples of possible open science workflows [20]



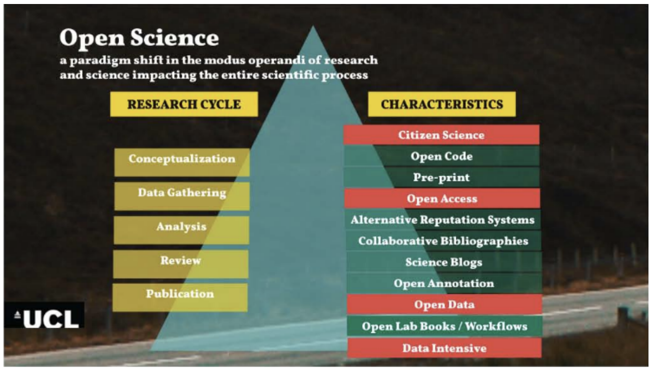
Adapted from Bosman and Kramer (2016): https://doi.org/10.6084/m9.figshare.3468641.v1

Figure 11. Research Workflow Cycles [52]

text of psychological science, structured into three phases: ‘Planning Your Research’, ‘Doing The Research’, and ‘Writing It Up’. The first consists in the pre-registration of a study report, which explicitly states study designs, hypotheses, and/or analysis plans, before the beginning of a research project, and its sharing through a repository. The second phase focuses on replicability and reproducibility, the first enabled by openly documenting the research process, sharing lab notebooks, full study protocols, and research materials, the latter by sharing documented data and code used for analysing them, even at intermediate stages of the research. The last phase consists in writing the manuscript and selecting a journal to publish in, as sharing a pre-print would be desirable. The workflow also suggests various tools that facilitate the creation and management of the research products, from version control systems (e.g., Git), for effectively enabling collaboration through the entire workflow, to research management tools (e.g., Open Science Framework), repositories (e.g., GitHub, figshare, Zenodo, PsyArXiv), format templates (e.g., papaja), and reference management software (e.g., Zotero).

W17 by Ayris and Ignat [5]

Ayris and Ignat [5] describe an Open Science workflow (Fig. 12) consisting in five phases, ‘conceptualization’, ‘data gathering’, ‘analysis’, ‘publication’, and ‘review’, characterised by ten horizontal dimensions, namely ‘citizen science’, ‘open code’, ‘open access’, ‘pre-prints’, ‘alternative reputation systems’, ‘science blogs’, ‘open annotation’, ‘open data’, ‘open lab books/workflows’, and ‘data-intensive approaches’.


Open Science

a paradigm shift in the modus operandi of research and science impacting the entire scientific process

RESEARCH CYCLE

Conceptualization
Data Gathering
Analysis
Review
Publication

CHARACTERISTICS

Citizen Science
Open Code
Pre-print
Open Access
Alternative Reputation Systems
Collaborative Bibliographies
Science Blogs
Open Annotation
Open Data
Open Lab Books / Workflows
Data Intensive

Figure 12. Open Science [5]

W18 by Klenk et al. [26]

Klenk et al. [26] show a transition from a traditional research paradigm to an Open Science one and the democratisation of the research lifecycle. The proposed lifecycle is contextualised in the biomedical field and consists of five steps: engagement or recruitment of individuals or populations, data generation, formulation of hypothesis, analysis, and publication. While the presented Open Science research approach does not actually change from a traditional one, data and methods are described in the publication and shared, fostering reuse and transparency. By sharing data and methods other researchers can discover and re-use them for generating additional hypotheses, conduct analysis, produce publications and share derivative data, so that their work can inform other researchers working on the shared assets.

W19 by Van Lissa et al. [55]

Van Lissa et al. [55] introduce the Workflow for Open Reproducible Code in Science (WORCS), a procedure that researchers can follow to facilitate the adoption of the Open Science paradigm, based on the first seven Transparency and Openness Promotion (TOP) guidelines [41]. It is implemented as an R package, but, since its conceptual foundations are platform independent, it can be used for deriving other solutions. The workflow is built upon free and open source software (e.g., Git, RStudio) and includes the use of the Open Science Framework for managing the entire research project. It consists of twenty steps divided into three phases: 'Study design', 'Writing and analysis', and 'Submission and publication'.

The first phase include six steps, of which three are optional: (1.1) "Create a (Public or Private) remote repository on a "Git" hosting service", (1.2) "When using R, initialize a new RStudio project using the WORCS template. Otherwise, clone the remote repository to your local project folder", (1.3) "Add a README.md file, explaining how users should interact with the project, and a LICENSE to explain users' rights and limit your liability. This is automated by the worcs package", (1.4) "Optional: Preregister your analysis by committing a plain-text preregistration and tag this commit with the label "preregistration"", (1.5) "Optional: Upload the preregistration to a dedicated preregistration server", and (1.6) "Optional: Add study materials to the repository".

The second phase consists of five steps: (2.1) "Create an executable script documenting the code required to load the raw data into a tabular format, and de-identify human subjects if applicable", (2.2) "Save the data into a plain-text tabular format like .csv. When using open data, commit this file to "Git". When using closed data, commit a checksum of the file, and a synthetic copy of the data", (2.3) "Write the manuscript using a dynamic document generation format, with code chunks to perform the analyses", (2.4) "Commit every small change to the "Git" repository", and (2.5) "Use comprehensive citation".

The remaining nine steps forms the third and last phase of the workflow: (3.1) "Use dependency management to make the computational environment fully reproducible", (3.2) "Optional: Add a WORCS-badge to your project's README file", (3.3) "Make a Private "Git" remote repository Public", (3.4) "Create a project page on the Open Science Framework (OSF) and connect it to the "Git" remote repository", (3.5) "Generate a Digital Object Identifier (DOI) for the OSF project", (3.6) "Add an open science statement to the Abstract or Author notes, which links to the "OSF" project page and/or the "Git" remote repository", (3.7) "Render the dynamic document to PDF", (3.8) "Optional: Publish the PDF as a preprint, and add it to the OSF project", and (3.9) "Submit the paper, and tag the commit of the submitted paper as a release of the submitted paper as a release, as in Step 4".

While fostering openness at every stage, the workflow is tailored to accommodate different needs, including various degrees of closeness, as the researcher can decide at what point the results are to be shared.

W20 by Wandl-Vogt et al. [57]

Wandl-Vogt et al. [57] describe a pilot-architecture of the open workflow for diversity4bio, a Linguistic diversity portal.

It consists of four phases, 'discover', 'explore', 'collect/share/publish', and 'invoke', enabled by the three layers that form the system architecture, namely 'Human interface Layer', 'Persistent Layer', and 'Enrichment Layer'. The first layer consists of the web interface and its components: 'Catalog', 'Visualization', and 'Private/Shared Workspace' supporting authentication. Through the catalogue the researcher can discover the resources, explore them with the visualisation component, and collect/share/publish them through the workspace. The first layer is based and enabled by the second one, which consists of a repository and a triple store. The last layer consists of the enrichment services that are invoked by a researcher to enrich/connect/link the data in the triplestore.

W21 by Bastille et al. [6]

Bastille et al. [6] describe a collaborative scientific workflow, defined in the context of the Integrated ecosystem assessment (IEA) framework and based on the scientific workflow described by Wickham and Grolemund [58] (Fig. 13).

It consists of five steps linked with different tasks: (*i*) 'Import', (*ii*) 'Analyze', (*iii*) 'Visualize', (*iv*) 'Communicate', and (*v*) 'Collaborate'.

In the first phase, data import is achieved using open science software tools, divided into three tasks, namely 'custom datasets', 'data catalogues', and 'data services' (e.g., ERDDAP, DataOne), the last two for finding and downloading data respectively.

Common software tools shared through GitHub or open source tools (e.g., Ecosim) can be used for the three tasks of the second phase, 'trend analysis', 'risk assessment', and 'ecosystem modelling'.

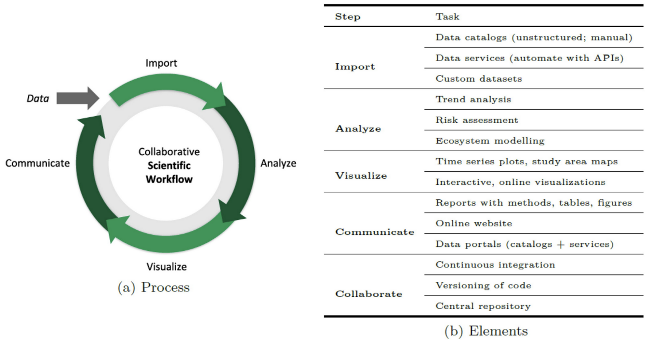
Visualisations, divided into the two tasks 'time series plots, study area maps' and 'interactive, online visualisations', are produced using open source tools (e.g., ggplot2 R library) and reproducible visualisation techniques in order to foster the development of standardised visualisations.

The communication of the findings includes three tasks, 'reports with methods, tables, figures' created through scripting-based reporting (using e.g., R Markdown, Jupyter Notebooks, Mathlab LiveScripts), 'online website' that allow interacting with data, and the creation of 'data portals, catalogs + services'.

The last step, 'collaborate', consists of the three tasks 'continuous integration' (CI), 'versioning of code', and 'central repository'. The use of CI services (e.g., Travis CI) allows the automatic update of reports and websites based on the latest available data. In order to enable collaboration, GitHub, built on the Git version control system, is used as a central open source code repository.

W22 by Pieper [45]

Pieper [45] presents a workflow for transparent and reproducible reporting on fee-based open access publishing using INTACT, a transparent infrastructure for open access publication fees based on an OLAP (Online Analytical Processing) Server and OLAP Cubes. It is illustrated in six steps: (*i*) the original data are submitted and (*ii*) shared on Github; (*iii*) data are automatically preprocessed in order to make them compatible with the enrichment scripts, enriched, and postprocessed (obtaining the final OpenAPC-compatible version) through



(a) Process

Step	Task
Import	Data catalogs (unstructured; manual)
	Data services (automate with APIs)
	Custom datasets
Analyze	Trend analysis
	Risk assessment
	Ecosystem modelling
Visualize	Time series plots, study area maps
	Interactive, online visualizations
Communicate	Reports with methods, tables, figures
	Online website
	Data portals (catalogs + services)
Collaborate	Continuous integration
	Versioning of code
	Central repository

(b) Elements

Figure 13. Scientific workflow within the IEA framework [6]

open source scripts and based on Crossref data; (*iv*) preprocessed data, preprocessing logs, scripts, enriched data, and postprocessed data are shared on GitHub; (*v*) postprocessed data are analysed through open source scripts; (*vi*) results are shared on GitHub. The local GitLab installation and GitHub are automatically synced.

W23 by Beck et al. [9]

Beck et al. [9] illustrate the open science workflow adopted in the context of the Tampa Bay Estuary Program Data Management Standard Operating Procedures (Fig. 14).

It consists of eight steps: (*i*) external data source; (*ii*) open-source toolkit; (*iii*) local data processing and document preparation; (*iv*) GitHub; (*v*) discoverable; (*v*) automated daily checks; (*vii*) interactive dashboard; and (*viii*) two pager report.

The workflow is based and built around GitHub and the open source tbeptools R package. In the first phase access to raw data is provided by project partners. The open source tools, shared through GitHub, are downloaded locally during the second phase and are used in the third phase for data processing and for creating the report. The tbeptools R package ensures that the data used locally are up to date by automatically comparing the local version with the one shared on GitHub. In the fourth phase the data (if they can not be found in the shared GitHub repository) and the results are uploaded to GitHub and made discoverable (fifth phase), while automated checks ensure that the uploaded data are synchronised to the latest available version (sixth phase). In the last two phases the results are shared through an interactive dashboard, a Shiny web application, and a pdf report, both hosted on the project site.

W24 by Assante et al. [4]

Assante et al. [4] present an example of open science workflow enabled by the D4Science Infrastructure and its Virtual Research Environments (VREs) (Fig. 15).

It consists of five phases: (*i*) 'Research Challenge', (*ii*) 'Data Collection', (*iii*) 'Analytics', (*iv*) 'Publication', and (*v*) 'Assessment'. In the 'Research Challenge' phase the social networking features of the VRE are used for informing the community about a research idea and/or creating a discussion around it. The VRE's workspace enables data sharing in the second phase. The data analytic platform of the VRE enables the third and homonymous phase. The fourth phase, the 'publication' phase, consists in the registration of an object (dataset, paper, etc.) in the shared catalogue, which is enabled by the VRE's publishing platform. The last phase, 'Assessment', is a vertical one, since the possibility of sharing every result in every phase enables the continuous evaluation of every research output through the entire workflow.

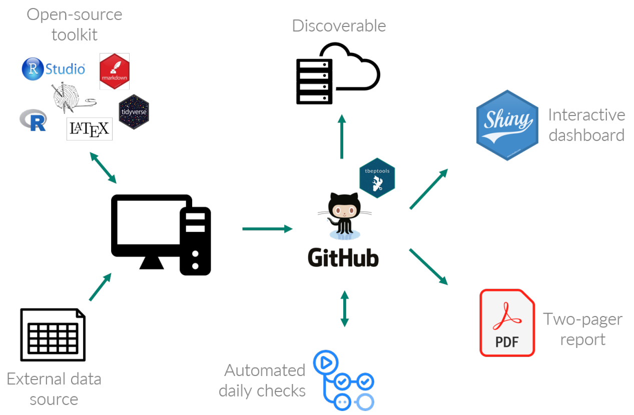


Figure 14. The TBEP open science workflow connecting source data to decision-support tools [9]

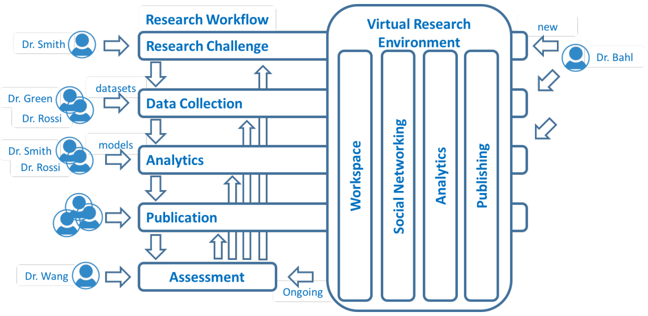


Figure 15. D4Science-enabled Workflow [4]

W25 by Grigorov et al. [19], Engineering National Academies of Sciences [14]

Grigorov et al. [19] present an open research lifecycle developed within the FP7 FOSTER training calendar 2014-2016, adapted from Tenopir et al. [51] (Fig. 16).

While it is not documented, nine stages are represented: (*i*) 'Idea & Proposal', (*ii*) 'Test & Method', (*iii*) 'Data & Observations', (*iv*) 'Model Code', (*v*) 'Research Articles', (*vi*) 'Review', (*vii*) 'Educate & Train', (*viii*) 'Policy Context', and (*ix*) 'Engage'. Every stage has associated the related open science processes, 'OS as part of concept', 'Open Notebook Science', 'RDM, Archive & Publish', 'Curate & Publish', 'Gold & Green OA', 'Open Peer-review', 'Open Educational

Resources', 'Policy Briefs & White Papers', and 'Science Literacy & Citizen Science' respectively.

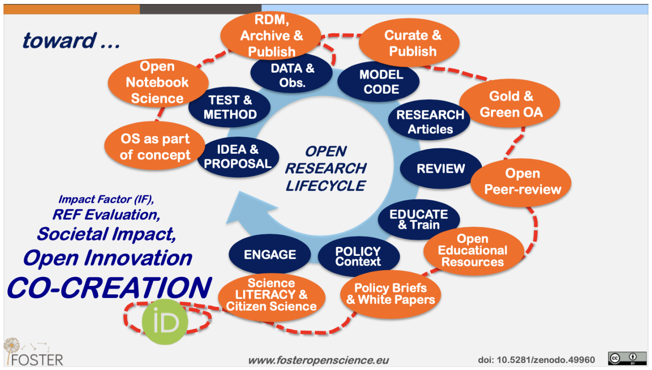


Figure 16. Open research lifecycle [19]

W26 by Gownaris et al. [18]

Gownaris et al. [18] illustrate a four-stage scientific life cycle defined in the context of mapping the barriers to the adoption of open science practices among early career researchers. Each stage is linked to the processes and concepts required for enabling open science practices.

In the first stage, 'Study Design & Tracking', open science is enabled by pre-registration and open processes, based on open notebook science and the related reproducibility of workflows, which allow access to every stage of the research, including negative results, and external scrutiny.

The second stage, 'Data Collection', is characterised by the use of open hardware ('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'), open infrastructure ('the sharing or giving access to the existing material culture and infrastructure of science'), open software ('source code must be freely available with a license – or terms and conditions – that allows for free dissemination and adaptation'), open data, and by citizen science.

In the third phase, 'Publication', open science is fostered by open access to every research product, open peer review (with its facets of open identities of the reviewers, public availability of the reviews, and open participation or crowdsourced peer-review), open data, and citizen science, in its acceptation of data source.

'Outreach' is the last phase of the life cycle, characterised by the adoption of open educational resources and by citizen science, in its acceptation of 'creating knowledge, sharing scientific skills and knowledge with the public, and promoting civic engagement in science'.

W27 by Kramer and Bosman [28], Labastida i Juan [31], Bosman and Kramer [10], Kramer and Bosman [29], Kramer and Bosman [30]

Kramer and Bosman [28] propose an example set of research practices that, during a series of workshops, which registered a wide participation of research stakeholders, have been identified as a possible part of an open science workflow.

The workflow consists of eight non-linear phases, articulated into 31 activities and 139 practices (Fig. 17).

The first phase, '*preparation*', encompasses three activities: 'project management/planning', 'crowdsource / define research priorities / ideas / collaborations', and 'fund / get contract'.

The second phase, '*discovery*', consists of five activities: 'search (literature / data / patents / code)', 'get alerts / recommendations', 'reference management', 'read / view', and 'annotate / tag (during / after reading)'.

The third phase, '*analysis*', consists of three activities: 'experiment & collect / mine / extract data', 'share notebooks / protocols / workflows', and 'analyze'.

The fourth phase, '*writing*', includes the activities 'write (+ code)', 'cite', and 'translate'.

The fifth phase, '*publication*', is articulated into the five activities: 'archive / share code', 'archive / share data (incl. video)', 'archive / share publication', 'select journal to submit to', and 'publish'.

The following and sixth phase, '*outreach*', consists of the activities 'archive / share posters', 'archive / share presentation', 'present research findings', 'outreach / valorization', and 'researcher profiling (& social network)'.

The last phase, '*assessment*', is composed by the activities 'peer review and commenting/recommending (pre-pub)', 'comment', 'peer review (post-pub)', 'measure impact (of output, e.g. article)', and 'assessment (of researcher / research group)'.

The '*preparation*' phase includes 16 practices: 4 in the 'project management / planning' activity ('giving everybody access to the needed infrastructure (even the wet lab)', 'managing projects openly', 'posting brief descriptions of ideas in a very early phase', and 'recording steps and inputs (reproducibility, credit, products)'); 8 in 'crowdsource / define research priorities / ideas / collaborations' ('crowdsourcing research topic prioritization', 'finding additional co-authors by early sharing of manuscripts', 'involving public / patients etc. in drafting research proposals', 'looking for research partners in the Global South', 'making expertise findable, accessible, visible & available when you need it', 'using immersive VR to enable widespread diverse collaboration', 'sharing your hypothesis before starting the data collection / analysis', and 'pre-registering studies'); and 4 in 'fund / get contract' ('crowdfunding (parts) of your research', 'openly publishing proposals', 'curating and sharing funding opportunities', and 'funding review, revision and improvement, not just novelty').

The '*discovery*' phase include 12 practices: 5 in the 'search (literature / data / patents / code)' activity ('improving findability by curating / tagging research objects', 'offering your service to Q&A platforms', 'extensively searching for existing data before generating your own', 'having open discovery of open access materials', and 'sharing your discovery process'); 'sharing your expert reading recommendations' in 'get alerts /

recommendations'; 3 in 'reference management' ('managing references collaboratively', 'sharing bookmarks / favourites', and 'sharing your collection of references'); 'share reading activities in order to make them usable as information filter' in 'read / view'; and 2 in 'annotate / tag (during / after reading)' ('tracking associations between 'annotations' and subsequent text changes / versions' and 'annotating papers, web pages').

'Analysis' include 14 practices: 4 in the 'experiment & collect / mine / extract data' ('making data & software immediately FAIR with PIDs as it is collected / generated', 'real time sharing of experiments through video', 'engaging in citizen science', and 'working with citizen science in academia'); 4 in 'share notebooks / protocols / workflows' ('sharing notebooks openly, online', 'sharing protocols openly, online', 'sharing workflows openly, online', and 'clarifying and specifically communicating materials and methods'); and 6 in 'analyze' ('sharing scripts of your analysis, openly online', 'getting help to check the reporting of your statistical analysis', 'discussing methodology/results early (e.g., on blogs)', 'documenting your analysis to allow full reproducibility', 'having open lab tests and so avoid being influenced by external parties', and 'using easily attainable (open source) software to allow anyone to reproduce your results').

'Writing' includes 13 practices: 7 in 'write (+ code)' ('writing collaboratively', 'drafting openly, online', 'using a collaborative authoring environment', 'making sure that all available info is also available to machines', 'publishing actionable papers including executable code and configurable visualizations', 'coding collaboratively', and 'having executable, forkable publications, including text, code and data'); 4 in 'cite' ('making citing 2-way: link back/track back', 'enabling "deep citing" at sentence/word level', 'citing OA versions of literature', and 'provide data/code citations'); and 2 in 'translate' ('spending money on translations' and 'translating research objects in world languages').

'Publication' includes 29 practices: 2 in the 'archive/share code' activity ('archiving & sharing code' and 'sharing executable scripts people can inject their own data in'); 6 in 'archive/share data (incl. video)' ('archiving & sharing data', 'storing data in the most open format possible', 'archiving & sharing video', 'archiving & sharing sound recordings', 'sharing all data when we can, explaining limitations when not possible', and 'sharing steps, packaging bits for reproduction'); 11 in 'archive / share publication' ('sharing all papers as preprints and calling these publications', 'depositing papers in a subject repository', 'depositing papers in an institutional repository', 'depositing papers in a preprint archive', 'archiving & sharing publications', 'archiving & sharing master/bachelor theses', 'archiving & sharing PhD theses', 'archiving & sharing book manuscripts', 'publish prepublication history (versions + peer reviews)', 'publishing pre-prints to encourage feedback / informal open peer review', and 'sharing research relationships as research objects'); 'selecting journal to submit to based on openness characteristics' in 'select journal to submit to'; and 9 in 'publish' ('making conflicts of interest transparant', 'specifying contributorship roles', 'using systematic versioning for publications', 'publishing Open Access', 'publishing Open Access in hybrid journals', 'declaring conflicts of interest in publications', 'using open licenses such as CC-BY or GNU-PL', 'flipping journals to become fully Open Access', and 'publishing papers in journals that judge only for rigour, not novelty').

The 'outreach' phase includes 29 practices: 'sharing posters online at same time as physical presentation' in the 'archive / share posters' activity; 'archiving & sharing presentations' in 'archive / share presentation'; 4 in 'present research findings' ('live blogging / tweeting from conferences etc.', 'doing live demo's at conferences', 'openly sharing your presentation slides etc. on day of presentation', and 'refusing to be part of all male or all white panels'); 18 in 'outreach/valorization' ('involving students, the lay public, and communicators in science communication', 'writing lay summaries', 'having clear communication of current research', 'incorporating practices of service into workflow (e.g., teaching, inclusivity)', 'having extra material (videos, slide ready figures, lay public figures, using Wikimedia)', 'making sure that research objects are understandable by all who need to understand them', 'writing plain language abstracts', 'presenting for the general public', 'writing for general magazines/newpapers', 'presenting and giving demo's for children (e.g., in primary/secondary education)', 'appearing on radio or television', 'creating dedicated outreach video', 'writing about (your) research in social media', 'reacting to messages in social media on your topic', 'using altmetrics for monitoring outreach', 'blogging and tweeting about every stage of the process', 'communicating analyzed data with experts, non-expert scientists and the lay-public', and 'learning from/engaging with communities'); and 5 in 'researcher profiling (& social network)' ('creating/maintaining an online researcher profile', 'using academic social networks to showcase your research output', 'using academic social networks to find and communicate with other researchers', 'using author IDs', and 'having a living knowledge network').

Under the last phase, 'assessment', are listed 25 practices: 7 in the 'peer review and commenting/recommending (pre-pub)' activity ('having all types of review openly available', 'sharing peer review reports openly', 'non-anonymous peer reviewing (names published)', 'using non-journal organized peer review', 'claiming credit for peer review', 'encouraging transparant peer review (& responses)', and 'testing reproducibility with peers as part of publishing process'); 2 in 'comment' ('commenting openly, online' and 'having an open, interoperable, transferable annotation layer over all scholarly objects'); 'writing/sharing post-pub peer reviews' in 'peer review (post-pub)'; 9 in 'measure impact (of output, e.g., article)' ('having accumulated research results usage data (as indicator of its impact)', 'visualizing how, why, what for, by whom research results were used', 'using narrative approaches to assess research', 'using openness data to assess research', 'using book level data to assess research', 'using usage / readership data to assess research', 'using metrics of commercial/social application to assess research', 'using altmetrics to assess research', and 'broadly considering all science when using parameters for calculating science'); and 6 in 'assessment (of researcher/research group)' ('using narrative approaches to assess researchers', 'using openness data to assess researchers', 'using book level data to assess researchers', 'using usage / readership data to assess researchers', 'using metrics of commercial / social application to assess researchers', and 'using altmetrics to assess researchers').

The 31st activity 'various', which is not part of the workflow's phases, includes the last practice 'making re-use and licensing/citation guidelines explicit'.

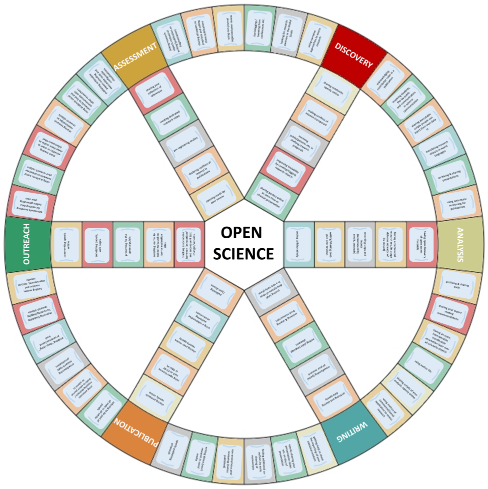


Figure 17. Wheel of Open Science practices [29]

W28 by Xiao [59]

Xiao [59] illustrates the shift from a research cycle towards a reproducible research cycle within the context of the scholarly communication infrastructure and services of the University of Hong Kong libraries.

The cycle is structured on the research products and consists of twelve steps, namely 'hypothesis', 'experimental design', 'raw data', 'processing / cleaning', 'code', 'tidy data', 'data curation service', 'open data', 'analysis', 'code', 'results', and 'article'. After the hypothesis formulation and the design of the experiment to verify it, raw data is collected and processed/cleaned into tidy data. Tidy data are then analysed, producing results that are then published in the form of an article. The code used for processing/cleaning the data and for analysing the tidy data is based on 'open research tools & platforms' and it is shared. The tidy data are curated by a data curation service and shared as open data through DataHub and accompanied by a data management plan (DMP) using a DMP platform. Each different research output (code, open data, and the article) can then concur to new forms of research impact and lead to new hypotheses.

W29 by Tse et al. [53]

While focusing on a list of available services, Tse et al. [53] describe also the application of open science practices in the context of COVID-19 vaccine / treatment development through a process composed of three phases: 'Research Activity', 'Raw Data', and 'Publication'. Each phase is linked to an open science approach: open source, open data, and open access respectively. The 'Research Activity' phase is in fact characterised by the use of software and services fostering collaboration, sharing and open and free availability. The 'Raw Data' phase is linked to platforms and open databases for sharing and re-use SARS-CoV-2-related data (e.g., Protein Data Bank). The last phase focuses on open access through preprint servers (bioRxiv, medRxiv) and open access journals (e.g., Public Library of Science journals).

W30 by Chávez Arroyo et al. [12]

Chávez Arroyo et al. [12] describe the application of the open-science philosophy in the context of creating a method for the simulation of Atmospheric Boundary Layer flows (Fig. 18).

The workflow is articulated into four stages: 'Open methodology', 'Open Source', 'Open Data', and 'Open Access'. With regard to the first stage, the methodology of the experiment is described and shared through open access journals and open source notebooks shared on GitHub, including the evaluation methodology of the benchmarks. About the second stage, 'Open Source', the model is itself open source and shared on GitHub, as well as the code used for the evaluations. In the third stage the evaluation data is open and shared on GitHub. In the last phase, as stated above, the method is submitted to open access journals.

Open methodology
Open Source
Open Data
Open Access

Figure 18. Open research lifecycle [12]

W31 by Minelli et al. [34, 35, 36, 37]

Minelli et al. [34, 35, 36, 37] present an open research lifecycle developed within the scope of oceanographic ecological research and based on Rüegg et al. [49].

The lifecycle is based on a spiral model (Fig. 19) and composed by fourteen sequential phases: (*i*) 'Plan', (*ii*) 'Share', (*iii*) 'Collect', (*iv*) 'Share', (*v*) 'Quality Assurance / Quality Control', (*vi*) 'Share', (*vii*) 'Analyze', (*viii*) 'Describe', (*ix*) 'Share', (*x*) 'Metadata', (*xi*) 'Share', (*xii*) 'Review', (*xiii*) 'Share', (*xiv*) 'Integration'. While every research object, from a research proposal to a paper, is to be shared as soon as it is produced, in particular this lifecycle is aimed to keep track of data, from their collection to their integration (phases iii to xiv).

W32 by Beck et al. [8]

Beck et al. [8] present an open science paradigm workflow,

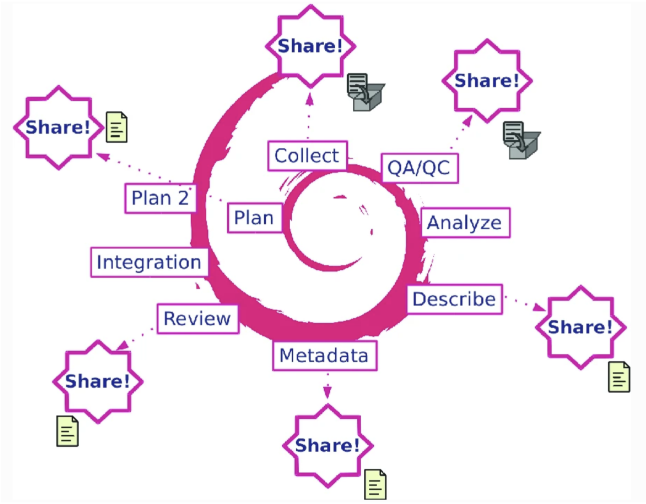


Figure 19. Open Research Lifecycle [34]

based on W12 by Hampton et al. [20], and its implementation developed in the context of creating products for the biological assessment of aquatic environments that can be effectively used by the environmental management community (Fig. 20).

The generic workflow consists of seven steps: (*i*) 'Conceptualize Project', (*ii*) 'Design, Collect & Analyze', (*iii*) 'Publish Open Data & Metadata', (*iv*) 'Publish Open Code', (*v*) 'Publish in Open Access', (*vi*) 'Inform Environmental Management', and (*vii*) 'Assess Environmental Response'. It is an iterative workflow since the availability, accessibility and openness of the research products at every stage foster collaboration and enable its application to every outcome.

Its implementation consists of twelve non-linear steps: (*i*) 'Identify research goals and pre-registration', (*ii*) 'Open planning', (*iii*) 'Managers and stakeholders', (*iv*) 'Collect and synthesize data', (*v*) 'Metadata', (*vi*) 'Data on open repository', (*vii*) 'External Data (RMP, NGO, academic)', (*viii*) 'Reproducible summary documents', (*ix*) 'Primary, Secondary literature', (*x*) 'Develop tool', (*xi*) 'Accessible tool (e.g., R package)', and (*xii*) 'Interactive applications'.

The stakeholders needs and the related research goals are identified through a two-way open planning with the management requiring the assessment product, enabled by online sharing of planning documents and by collaboration and communication tools (e.g., Google documents, Slack). The resulting study design is pre-registered through e.g., the Open Science Framework (`https://osf.io/`) or AsPredicted (`https://aspredicted.org/`). Following the study design and pre-registration, data sources are identified and synthesized data products are created, documented, and curated using a metadata standard (e.g., Ecological Metadata Language) and e.g., Zenodo to assign a DOI. The synthesized data are then deposited in an open repository and possibly used for creating interactive applications and automatically generated summary reports (by using e.g., knitr [60], RMarkdown [7], Jupyter notebooks [27]), which can effectively convey the assessment information to managers and stakeholders. These research products, once shared, can in turn contribute directly to further scientific studies and advances. The tools for creating the interactive applications and generating the reports can be developed using R and openly shared by the research team or can be reused from existing projects and R packages (e.g., shiny for creating interactive applications, the raster package, the bioassessment specific package TITAN2).

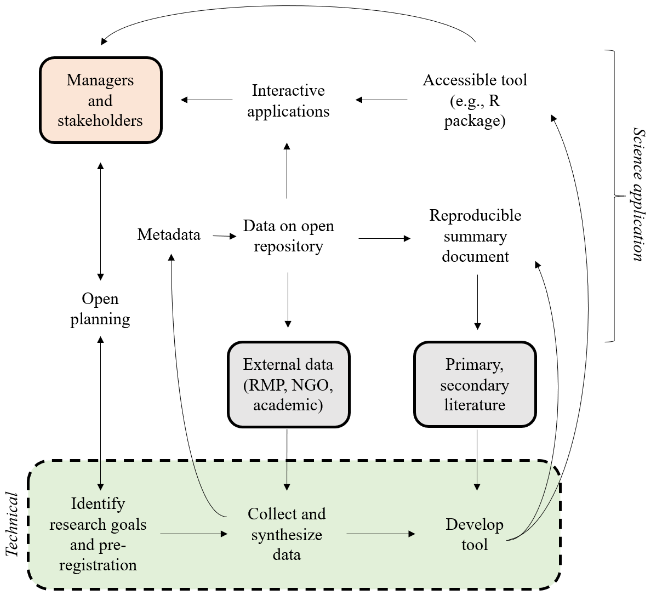

Managers and stakeholders
Interactive applications
Accessible tool (e.g., R package)
Science application
Metadata
Data on open repository
Reproducible summary document
Open planning
External data (RMP, NGO, academic)
Primary, secondary literature
Technical
Identify research goals and pre-registration
Collect and synthesize data
Develop tool

Figure 20. Open Research Lifecycle [8]

W33 by Reimer et al. [48]

Reimer et al. [48] discusses the inclusion of open science practices in the scientific workflow of early-career researchers at the Control of Impulsive Action (Ctrl-ImpAct) Lab and defines the open science workflow named 'Co-Pilot' system.

The system is structured into three phases, 'Study preparation & Hypotheses', 'Data Collection & Analysis', and 'Manuscript & Communication', each linked to an open science aspect, open methodology, open data & open source, and open access respectively.

In the first phase the study is pre-registered using the Open Science Framework or via registered reports, specifying the motivating research question and hypothesis, the research design and study materials including planned sample size, the outcome variables, and the predictor variables and a more specific data analysis plan, before data collection has started.

In the second phase a data management plan is created using DMPonline.be. Once a project is completed, data are made FAIR compliant and anonymised data are deposited through the Open Science Framework or a trusted institutional repository under a CC BY 4.0 licence. Data is also accompanied by related documentation that facilitate its interpretation. The scripts used for analysing the data are realised using open source software (e.g., PsychoPy, jsPsych, R) and shared through the Open Science Framework under the GNU General Public License (GPL) 3.0 with the relevant study material under a CC BY 4.0 licence.

In the third and last phase pre-prints are shared through e.g., psyarxiv.com while the final manuscript is published in fair/nonprofit Open Access journals via Diamond or Gold Open Access. Green Open Access is considered only if the first two are not an option.