Research Software vs. Research Data II: Protocols for Research Data dissemination and evaluation in the Open Science context

Teresa Gomez-Diaz, Tomas Recio

RESEARCH ARTICLE

REVISED Research Software vs. Research Data II: Protocols for Research Data dissemination and evaluation in the Open Science context [version 2; peer review: 2 approved]

Teresa Gomez-Diaz[1], Tomas Recio[2]

[1]Laboratoire d'informatique Gaspard-Monge, CNRS, Paris-Est, France
[2]Universidad Antonio de Nebrija, Madrid, Spain

v2 **First published:** 28 Jan 2022, **11**:117
https://doi.org/10.12688/f1000research.78459.1
Latest published: 07 Oct 2022, **11**:117
https://doi.org/10.12688/f1000research.78459.2

Abstract

Background: Open Science seeks to render research outputs visible, accessible and reusable. In this context, Research Data and Research Software sharing and dissemination issues provide real challenges to the scientific community, as consequence of recent progress in political, legal and funding requirements.
Methods: We take advantage from the approach we have developed in a precedent publication, in which we have highlighted the similarities between the Research Data and Research Software definitions.
Results: The similarities between Research Data and Research Software definitions can be extended to propose protocols for Research Data dissemination and evaluation derived from those already proposed for Research Software dissemination and evaluation. We also analyze FAIR principles for these outputs.
Conclusions: Our proposals here provide concrete instructions for Research Data and Research Software producers to make them more findable and accessible, as well as arguments to choose suitable dissemination platforms to complete the FAIR framework. Future work could analyze the potential extension of this parallelism to other kinds of research outputs that are disseminated under similar conditions to those of Research Data and Research Software, that is, without widely accepted publication procedures involving editors or other external actors and where the dissemination is usually restricted through the hands of the production team.

Keywords

Research Data, Research Software, Open Science, Research outputs' dissemination, Research Evaluation, FAIR principles.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 07 Oct 2022	✓ view	✓ view
version 1 28 Jan 2022	? view	? view

1. **Charles Romain**, Imperial College London, London, UK
Henry S. Rzepa, Imperial College London, London, UK

2. **Mark Leggott**, Digital Research Alliance of Canada, Ottawa, Canada

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Research on Research, Policy & Culture gateway.

Corresponding authors: Teresa Gomez-Diaz (teresa.gomez-diaz@univ-mlv.fr), Tomas Recio (trecio@nebrija.es)

Author roles: Gomez-Diaz T: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Recio T** : Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work is partially funded by the CNRS-International Emerging Action (IEA) PREOSI (2021-22).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Gomez-Diaz T and Recio T. **Research Software vs. Research Data II: Protocols for Research Data dissemination and evaluation in the Open Science context [version 2; peer review: 2 approved]** F1000Research 2022, **11**:117 https://doi.org/10.12688/f1000research.78459.2

First published: 28 Jan 2022, **11**:117 https://doi.org/10.12688/f1000research.78459.1

REVISED **Amendments from Version 1**

This version considers the comments of the reviewers to better explain and illustrate some of the concepts presented in the article.

In particular we have improved Sections 2.4 and 3.4 to better explain the relations between the FAIR Principles and the work we propose here. A new table is added to Section 3.4 to show the connections between the RD CDUR evaluation protocol of Section 3.3 and the FAIR principles of Reference 1.

See our answers to the referee reports to complete the differences with the version 1 of this article.

Any further responses from the reviewers can be found at the end of the article

1. Introduction

Researchers produce many different outputs in their work in order to obtain the results that will be published in scientific journals, in articles that are still the main exchanging information mechanism in the scientific conversation. Among others, researchers produce Research Data (RD) and Research Software (RS), but yet again, both outputs have not currently a publication procedure as widely accepted as the one existing for articles, which constitutes one of the main drawbacks for their acceptance as *first class citizens* in the scientific ecosystem. This is one of the goals of the FAIR guiding principles 1:

> *… is for scholarly digital objects of all kinds to become 'first class citizens' in the scientific publication ecosystem, where the quality of the publication – and more importantly, the impact of the publication – is a function of its ability to be accurately and appropriately found, reused, and cited over time, by all stakeholders, both human and mechanical.*
>
> *[…] we do not pay our valuable digital objects the careful attention they deserve when we create and preserve them.*

On the other hand, the following definition sets up Open Science goals related to research outputs 2:

> *Open Science is the political and legal framework where research outputs are shared and disseminated in order to be rendered visible, accessible and reusable.*

In this context, as reported in 3, the necessary skills to reach out these goals are complex:

> *The skills needed for Open Science cover a broad span from data management to legal aspects, and include also more technical skills, such as data stewardship, data protection, scholarly communication and dissemination (including creating metadata) …*

and still require to be engineered 4 (see also section 5 of 5):

> *An acceptable workflow needs to be created. However, most researchers, while experts in their own fields, have little awareness of metadata standards for data publication and information science in general, leading to cognitive and skill barriers that prevent them from undertaking routine best-practice data management.*

Another drawback of this missing publication procedure for RD and RS is the possible loss of the expert knowledge that has been acquired along the research process 6:

> *If not traditional papers and volumes, what, then, should researchers be publishing? Whilst the digital exchange of data is straightforward, the digital exchange and transfer of scientific knowledge in collaborative environments has proven to be a non-trivial task, requiring tacit, and rapidly changing expert knowledge – much of which is lost in traditional methods of publication and information exchange. We believe that there is a need for mechanisms that support the production of self-contained units of knowledge and that facilitate the publication, sharing and reuse of such entities.*

Examples of this lost knowledge include the report of failure cases, which are rarely published; or the description of the modifications that have been included in the final implemented algorithms, and that are the result of a long trial and error process to improve the initially conceived algorithm or to avoid computational errors.

Although the current trend in the scientific publication ecosystem is to place RS and RD into a better position, many researchers are still at a loss when facing RS and RD dissemination, and do not possess the needed skills, support or assistance for their disclosure in the right conditions. Moreover, they consider that much work and effort would be necessary to accomplish this goal, while having little or no positive effect in their curriculum 4:

> *Put crudely, the large amount of effort involved in preparing data for publication release, coupled with the negligible current incentives and rewards, prevents many researchers from doing so.*

Notice that some simple rules to be considered in RD dissemination practices have been already proposed, for example, in 7

On the other hand, research funders, like the European Commission, are currently laying out Open Science policies in their calls, in which it is required open access to the generated RD of the funded projects (although there may be exceptions), and where it is recommended to provide open access to research outputs in all generality, beyond publications and data, e.g. software tools 8. Notice that in the dissemination of these research outputs it is necessary to provide significant information in order to facilitate their visibility, accessibility and their reuse 9:

> *Detailed provenance includes facets such as how the resource was generated, why it was generated, by whom, under what conditions, using what starting-data or source-resource, using what funding/resources, who owns the data, who should be given credit, and any filters or cleansing processes that have been applied post-generation.*

Bearing in mind the above described landscape, the goal of our work here is to contribute to the improvement of the scientific endeavor with protocols that could help researchers, and the community at large, in the dissemination of their produced RD and RS, while contributing to the accomplishment of Open Science goals.

We concentrate here in practical matters, that is, in the *how to*: how to disseminate RD and RS to make them *first class citizens* so that they become visible, accessible, reusable. But dissemination procedures are not enough. With the aim to motivate researchers to deal with better dissemination tasks, most of the times considered by the members of the scientific community as an additional, useless burden, we should also take into consideration pathways that yield improved research evaluation practices, so relevant for researchers. That is, pathways that contribute to evaluate correctly the disseminated outputs with protocols that help both the researchers – to know what will be evaluated and how – as well as the evaluators – into setting the evaluation process.

Our proposal is grounded on our knowledge and experience concerning RS 10-14. This translation of knowledge from RS to RD has been already successfully applied 15 to propose a RD definition and to tackle the Borgman's conundrum challenges 16. In the present paper we attempt to extend this approach to the case of RD dissemination and evaluation practices. Indeed, there are some obvious differences between software and data. But by keeping these aspects aside, and focusing on the similarities, we can learn a lot from the common features that appear in the production context of RD and RS. As remarked above, this is a general procedure that can be adapted to several situations. Even when the differences are too important, and maybe the proposed dissemination and evaluation procedures are not directly applicable as such in both settings, they could help to suggest hints to address the diverse issues appearing in each environment. In summary, the present work follows and expands the approach adopted in 15. Both articles can be read separately, but they constitute a whole.

The plan of this work is as follows. The next section is devoted to revisit the corresponding points related to RS: definition, dissemination, evaluation and consideration of the role of FAIR principles in this context. Section 3 focus then in RD topics, reviewing the proposed RD definition 15 and to present the main contribution: some comprehensive RD dissemination and evaluation procedures. Conclusions will end this work.

2. Research Software

Three are the main components of this section: the RS definition coming from 13, 14, the RS dissemination procedure coming from 10, the CDUR RS evaluation protocol from 13. Some comments on FAIR principles for RS will complete this section.

2.1. Research Software definition, reference and citation

In this work we consider the following definition of RS 13, 14:

> ***Research software*** *is a well identified set of code that has been written by a (again, well identified) research team. It is software that has been built and used to produce a result published or disseminated in some article or*

> *scientific contribution. Each research software encloses a set (of files) that contains the source code and the compiled code. It can also include other elements as the documentation, specifications, use cases, a test suite, examples of input data and corresponding output data, and even preparatory material.*

We observe, following the above definition, that RS has three main characteristics:

- the goal of the RS development is to do research,
- it has been written by a research team,
- the RS is involved in the obtention of the results presented in scientific articles (as the most important means for scientific exchange are still articles published in scientific journals), or by any other kind of recognized scientific means.

Note that documentation, licenses, examples, data, tests, software management plans and other related information and materials can also be part of the set of files that constitutes a specific RS.

Moreover, a RS development team may not just use software produced by other teams, but also include external software as a component inside the ongoing development, something which can be facilitated by the Free/Open Source Software (FLOSS) [1] licenses. This potential external component will qualify here as RS if it complies with the three characteristics given in the above definition 15. Moreover, the responsible team of the resulting work should clearly identify the included external components and their licenses, as well as highlight, by means of recommended citation practices, 13, 17, 18, the external components that qualify as RS.

General aspects of FLOSS issues can be consulted, for example, in 19. Let us remark that good practices for software development management ask for updating regularly the RS related information, like, for example, project's funding, publications or involved teams and contributors. A Software Management Plan (SMP) can be a powerful tool to help and to handle this information, see for example 12, and the references therein.

Let us recall that RS reference and citation recommendations have been considered in section 2.5 of 13 where we propose easy to adopt methods to improve RS citation practices, other citation related works can be found in 17, 18, 20.

2.2 A Research Software dissemination procedure

Let us begin by recalling that, as stated in 8:

> *Dissemination means the public disclosure of the results by appropriate means (other than resulting from protecting or exploiting the results), including by scientific publications in any medium.*

The following RS dissemination procedure has been proposed in 10 and was first published[2] in the PLUME project[3] (2006-2013) 13, 21. The French initial version includes a close analysis of legal issues (French author rights, licensing) in order to produce FLOSS RS. It is slightly updated and completed in the following. More information on the legal issues can be found in 11, or in section 2.1 of Reference 15.

As a general recommendation, it is best practice to consider licensing issues and to keep the related information in a SMP from the very first stages of the RS development. The RS license establishes its sharing conditions: it can give rights for access, copy, modification, redistribution of the RS, and it can establish reciprocity clauses that should be respected by the potential RS users. Licenses should be put well into place before releasing the RS.

Here we present the proposed RS dissemination procedure. Steps marked with (*) are to be revisited regularly for each version release.

[1] https://en.wikipedia.org/wiki/Free_and_open-source_software
[2] Diffuser un logiciel de laboratoire: recommandations juridiques et administratives, 2010, https://zenodo.org/record/7096216. In French.
[3] The PLUME project platform has been closed in 2022. Some of the documents published in this platform are now available from https://zenodo.org/communities/plume-patrimoine-logiciel-laboratoire/

- Choose a name or title to identify the RS, avoid trademarks and other proprietary names, you can associate date, version number, and target platform. Consider best practices in file names[4].

- (*) Establish the list of authors and affiliations (this is the so called *research team step*). An associated percentage of participation, completed with minor contributors can be useful. If the list is too long, keep updated information in a web page or another document like a SMP, for example, where you can mention the different contributor roles. This is the step in which the intellectual property producer's rights are established. Producers include the RS authors and rightholders. This is then the step in which RS legal issues related to copyright information are dealt with.

- (*) Establish the list of included software and data components, indicate their licenses (or other documents like the component's documentation) giving the rights to access, copying, modification and redistribution for each component. In the case of software and data that fall in the category of RS or RD, please take into consideration best citation practices 13, 17, 18.

- Choose a software license, with the agreement of all the rightholders and authors, and establish a signed agreement if possible. The licenses of the software components that have been included and/or modified to produce the RS can have impact in your license decision, see for example 10, 19, 22. Software licenses and licensing information can be found at the Free Software Foundation (FSF)[5], the Open Source Initiative (OSI)[6], and the Software Package Data Exchange (SPDX)[7]. Consider using FLOSS licenses to give the rights of use, copy, modification, and/or redistribution. This is then the step in which legal issues related to the RS sharing conditions are to be taken into consideration. Indicate the license in the RS files, its documentation, and the project web pages. Give licenses, like GNU FDL[8], Creative Commons (CC)[9], LAL[10], to documentation and to web sites.

- Choose a web site, forge, or deposit to distribute your product; licensing and/or conditions of use, copy, modification, and/or redistribution should be clearly stated, as well as the best way to cite your work. Good metadata and respect of open standards are always important when giving away new components to a large community: it helps others to reuse your work and increases its longevity. Use Persistent Identifiers (PIDs)[11] if possible.

- (*) This step deals with the utility of the RS and how it has been used for your research (this is the *research work step*). Establish the list of main functionalities, and archive a tar.gz or similar for the main RS versions in safe place. Keep a list of the associated research work, including published articles. Update your documentation, SMP, web site, etc. with the new information in each main RS version.

- Inform your laboratories and head institutions about this RS dissemination (if this has not be done in the license step).

- Create and indicate clearly an address of contact.

- Release the RS.

- Inform the community (e.g via mailing lists), consider the publication of a software paper, see for example the list of Journals where you can publish articles focusing on software[12].

---

[4]See for example https://libguides.princeton.edu/c.php?g=102546&p=930626, https://doranum.fr/stockage-archivage/comment-nommer-fichiers/
[5]https://www.fsf.org/licensing/
[6]https://opensource.org/licenses
[7]https://spdx.org/licenses/
[8]http://www.gnu.org/copyleft/fdl.html
[9]https://creativecommons.org/choose/
[10]http://artlibre.org/licence/lal/en/
[11]http://en.wikipedia.org/wiki/Persistent_identifier
[12]This list is mantained by Neil Chue Hong in the Software Sustainability Institute web page https://www.software.ac.uk/which-journals-should-i-publish-my-software

This proposed procedure is flexible and can be adapted to many different situations. It has been taken into consideration in the HAL research software deposit 23[13].

2.3 The CDUR procedure to evaluate Research Software

We include in this section the summarized version of the CDUR protocol that can be found in 13 (section 4.1). This reference gives a detailed description and analysis of the protocol as well as a complete list of references related to this work. This procedure for RS evaluation contains four steps to be applied in the following chronological order: **C**itation, **D**issemination, **U**se and **R**esearch. For example, as we have seen in the Section 2.2, the first steps in the RS dissemination procedure correspond to the correct RS identification, and in order to be correctly cited, the RS reference should be clearly indicated. Let us introduce a resumed version of these four steps.

(C) Citation. This step measures to what extent the evaluated RS is well identified as a research output. It is also the step where RS authors are correctly identified as well as their affiliations.

Section 2.5 of 13 proposes three different ways to establish a RS reference, in order to facilitate its citation, formula that can include the use of persistent identifiers. Moreover, a more evolved RS identification level could be provided in the form of a metadata set. Reference and metadata include, among other informations, the list of the RS authors and their affiliations (13, section 2.2). See also 17, 18, 20.

(D) Dissemination. This step measures the quality of the RS dissemination plan involving actions such as:

- Choosing a license, with the agreement of all the rights' holders and authors. Consider, preferably, using FLOSS licenses.
- Choosing a web site, forge, or deposit to distribute the product; stating clearly licensing and conditions of use, copy, modification, and/or redistribution.
- Creating and indicating a contact address.

This step deals with legal issues involving the authors and rightholders (as established in the **C**itation step) deciding and installing the license(s) for the RS dissemination. This is also the step concerning Open Science, as the RS license expresses its sharing conditions; and where policy makers should establish the Open Science policies that will be applied in the evaluation process.

Finally, let us recall that the inclusion of the list of related publications, data sets and other related works in the dissemination procedure helps to prepare the reproducible science issues that are to be taken into account in the **U**se step.

(U) Use. This step is devoted to the evaluation of the technical software aspects. In particular, this step measures the quality of the RS usage, considering that a performing RS is one that is both correct and usable by the target scientific community.

The RS usability does not only refer to the quality of the scientific output but also can deal with other matters, such as the provided documentation, tutorials and examples (including both inputs and outputs), an easy and intuitive manipulation, testing and version management, etc.

This is the reproducible science step, where it is measured how the published results obtained with the RS can be replicated and reproduced.

(R) Research. This step measures the impact of the scientific research that has required in an essential way the RS under consideration.

The evaluation of this item should follow whatever standards for scientific research quality in the concerned community.

[13]As remarked in a previous note, the *documents de référence PLUME pour mieux gérer les développements logiciels, les diffuser et les valoriser dans un laboratoire* https://zenodo.org/communities/plume-patrimoine-logiciel-laboratoire/

This is the step where the RS related publications (as described in the RS definition in Section 2.1) come into play, and where the evaluation should consider the difficulty of the addressed scientific problems, the quality of the obtained results, the efficiency of the proposed algorithms and data structures, etc. The RS impact can also be assessed through the research impact of the related publications, and through its inclusion (or use) as software component in other RS.

Each of these four steps can reach different levels of qualification and the corresponding scale is to be set up by the policy makers considering a particular evaluation event. Thus, the CDUR protocol can be easily adapted to different circumstances: career evolution, recruitment, funding, RS peer review or other procedures to be applied by universities and other research performing institutions, research funders, or scientific journals, and it can also be adapted to different evaluation situations arising in different scientific areas.

2.4 FAIR Research Software

Although the FAIR principles have been first designed for data, they apply as well to other digital objects 1:

> *… it is our intent that the principles apply not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects – from data to analytical pipelines – benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.*

In Reference 24 we can find further explanation on the FAIR Guiding Principles[14]:

> *FAIR refers to a set of principles, focused on ensuring that research objects are reusable, and actually will be reused, and so become as valuable as is possible. They deliberately do not specify technical requirements, but are a set of guiding principles that provide for a continuum of increasing reusability, via many different implementations. They describe characteristics and aspirations for systems and services to support the creation of valuable research outputs that could then be rigorously evaluated and extensively reused, with appropriate credit, to the benefit of both creator and user.*

Explanation that is followed by a list of items to outline what FAIR is not. We complete here this list of what FAIR is not with the following: FAIR does neither claim to be, strictly speaking, a dissemination procedure, nor an evaluation protocol as the ones proposed in the present article. Yet these FAIR Guiding Principles give instructions that can be considered in dissemination procedures and evaluation protocols.

In the case of RS, FAIR principles have been considered in several conferences and publications, although some adaptations seem to be necessary 20, 25, 26. See also the documents in the FAIR Research Software (FAIR4RS) Zenodo Community[15] of the RDA FAIR4RS WG[16] 27.

In this section we highlight two points regarding these principles that appear in our RS dissemination procedure (see Section 2.2) and the CDUR evaluation protocol (see Section 2.3), namely those referring to Persistent Identifiers (PIDs) and metadata, as remarked in 5:

> *Central to the realization of FAIR are FAIR Digital Objects. These objects could represent data, software, protocols or other research resources. They need to be accompanied by Persistent Identifiers (PIDs) and metadata rich enough to enable them to be reliably found, used and cited.*

Note that these two points are included in the basic "minimum standard" of 5 (p. 13). In particular we would like to observe the following points regarding PIDs:

- we recommend to use PIDs associated to authors, like ORCID[17],
- we recommend to associate PIDs to the disseminated RS; as a RS can have several versions, do consider a different PID for each main release,

[14] See also https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/
[15] https://zenodo.org/communities/fair4rs/
[16] https://www.rd-alliance.org/groups/fair-research-software-fair4rs-wg
[17] https://orcid.org/

- as PIDs can be provided by the chosen deposit, PID provision should be one of the arguments favoring the selection of a deposit like, for example, Zenodo[18],

- articles associated to the RS should have their own PID, furnishing in this way the RS with other possible citation forms 13 (section 2.5), i.e. with complementary means to reliably finding the RS, facilitating thus its use and citation by other researchers,

- if the included data and software components or other external components that are necessary to run the disseminated RS have associated their own PIDs, it is convenient to refer to them in order to contribute to their own access and visibility.

Note that several possibilities for software identification are discussed in 27.

On the other hand, concerning the role of metadata sets in our RS dissemination and evaluation proposals, let us observe that metadata is a very flexible concept, going from a simple reference or citation form or the use of citation file formats 17, to a very complete and precise RS description. In any case, our protocols consider that they are an important tool to set attribution to the RS and to facilitate credit. One possibility we would like to suggest is the metadata format proposed in the PRESOFT SMP template 12, that has a manageable size and has also the advantage that it is based in the RS index card elaborated in the PLUME project (2006-2013). A different, more complex metadata set can be generated, for example, with COdeMeta[19]. We remark that it is the role of the RS producer team to set the RS metadata complying with FAIR principles 1, 27, that appear in the Findable, Interoperable and Reusable guidelines, and to ensure that the RS deposit guarantees that the metadata remain accessible (principle A.2). These aspects related to metadata and citation forms are to be considered in the **C,** and **U** steps of our CDUR protocol following the requirements established by the evaluation committees.

Finally, we consider the implementation and adoption of FAIR principles 1, 5, 9, 27, 29 and other standards as arguments favoring the choice of a deposit for the RS, see the principles F4 and the four Accessible ones (A.1, A1.1, A1.2, A2) of 1 and 27, which are to be considered in the D step of our CDUR protocol. A tool to help taking such decision could be the FAIRsharing platform[20], that provides a large amount of community-developed standards, as well as indicators (among others) necessary to monitor their adoption, and to follow data policies established by funders, editorials and other organizations. See, for example, the information that appears in the FAIRsharing platform associated to the FAIR Principles[21].

A table establishing relationships between the CDUR evaluation protocol steps and the FAIR principles has been included in Section 3.4. We refer the reader to this table and the subsequent comments for further comparison issues.

3. Research Data

This section translates to RD the previously addressed RS issues: definition, dissemination and evaluation, ending with some RD FAIR considerations.

3.1 A Research Data definition

In coherence with the declared parallelism between RS and RD, we consider here the RD definition proposed in 15.

> ***Research Data*** *is a well identified set of data that has been produced (collected, processed, analyzed, shared and disseminated) by a (again, well identified) research team. The data has been collected, processed and analyzed to produce a result published or disseminated in some article or scientific contribution. Each research data encloses a set (of files) that contains the dataset maybe organized as a database, and it can also include other elements as the documentation, specifications, use cases, and any other useful material as provenance information, instrument information, etc. It can include the research software that has been developed to manipulate the dataset (from short scripts to research software of larger size) or give the references to the software that is necessary to manipulate the data (developed or not in an academic context).*

[18]https://zenodo.org
[19]https://codemeta.github.io/codemeta-generator/
[20]https://fairsharing.org/
[21]https://fairsharing.org/FAIRsharing.WWI10U

Thus, as carefully argued and commented in 15, RD has three main characteristics:

- the goal of the RD collection and analysis is to do research, that is, to answer a scientific question,
- it has been produced by a research team,
- the RD is involved in the obtention of the results presented in scientific articles (as the most important means for scientific exchange are still articles published in scientific journals), or by any other kind of recognized scientific means.

The reader is referred to 15 for a thorough explanation of the different issues involved in the description of these characteristics. Yet, we recall in here that the identified set of data constitute a database in the case the data are arranged in a systematic or methodical way and is individually accessible by electronic or other means 30-34. The *sui generis* database rights primarily protects the producer of the database and may prohibit the extraction and/or reuse of all or a substantial part of its content for example 30.

Remark that it is becoming a general practice for research funders to ask for a Data Management Plan (DMP) concerning the data generated in a funded project[22] 8, 37-38. See for example the DMPonline platform of the Digital Curation Center (DCC) as a helpful tool to create, review, and share DMPs that meet institutional and funder requirements[23]. In particular, French research projects can benefit from DMP OPIDoR[24].

3.2 A procedure for Research Data dissemination

The following procedure has been adapted to RD from the RS dissemination procedure proposed in Section 2.2. Only a new item has been added here for RD (the 3rd item) to highlight the potential difficulties concerning legal (and ethical) issues. Similarly, steps marked with (*) are to be revisited regularly in each version release, if necessary.

Again, as a general recommendation, it is best practice to consider licensing issues 34 and to keep a DMP from the very first stages of the RD development. The RD license establishes the sharing conditions: it can give rights for access, copy, modification, redistribution of the RD, and it can establish reciprocity clauses that should be respected by the potential RD users. It should be put well into place before releasing the RD.

- Choose a name or title to identify the RD, avoid trademarks and other proprietary names, you can associate date, version number … Consider best practices in file names[25].
- (*) Establish the list of the persons that have participate to the RD production, that is, the persons who have collected, processed, analyzed, shared and disseminated the RD; as well as their affiliations (this is the so called *research team step*). If the list is too long, keep updated information in a web page or another document like a DMP, for example, where you can mention the different contributor roles. This is the step in which the producer's rights are established, if any. Producers include the RD authors (in the case there are intellectual property rights associated to the RD) and the corresponding rightholders. This is then the step in which legal issues related to copyright and ownership information are dealt with 32, 33, 37, 38.
- Data can have associated other legal (or ethical) contexts 15, 34, 38, 39, they can be intimately related to the ongoing research work, consider them with the help of legal experts if necessary.
- (*) Establish the list of included software and data components, indicate their licenses (or other documents like the component's documentation) giving rights to access, copying, modification and redistribution for the component. In the case of software and data that fall in the category of RS or RD, please take into consideration best citation practices 40–43, 13, 18, 20

[22]https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
[23]https://dmponline.dcc.ac.uk/
[24]https://opidor.fr/planifier/
[25]See for example https://libguides.princeton.edu/c.php?g=102546&p=930626, https://doranum.fr/stockage-archivage/comment-nommer-fichiers/

- Choose a data license, with the agreement of all the producers and rightholders, and establish a signed agreement if possible. The licenses of data components that have been included and/or modified to produce the RD can have impact in your license decision 34. Consider using licenses like the Creative Common licenses (V4.0) [26] or the Open Data Commons Licenses[27], for example. Other data licenses can be found at SPDX[28]. This is then the step in which legal issues related to the RD sharing conditions are to be taken into consideration. Indicate the license in the RD files, its documentation, the project web pages, etc. Give licenses, like GNU FDL[29], Creative Commons (CC)[30], LAL[31], to documentation and to web sites.

- Choose a web site, forge, a data repository or any another deposit to distribute your product; licensing and conditions of use, copy, modification, and/or redistribution should be clearly stated, as well as the best way to cite your work. Good metadata and respect of open standards are always important when giving away new components to a large community: it helps others to reuse your work and increases its longevity. Use Persistent Identifiers (PIDs) [32] if possible. This point corresponds to the Where? question of the Borgman's conundrum challenges as discussed in the Conclusions section of 15.

- (*) This step deals with the utility of the RD and how it has been used for your research (it is then the *research work step*). Establish the list of the main RD research issues that appear in your work and that can facilitate its reuse. Archive a tar.gz or similar for the main RD versions in safe place. Keep a list of the associated research work, including published articles. Update your documentation, DMP, web site … with the new information in each main release.

- Inform your laboratories and head institutions about this RD dissemination (if this has not be done in the license step).

- Create and indicate clearly an address of contact.

- Release the RD.

- Inform the community (e.g. via mailing lists), consider the publication of a data paper.

This proposed procedure is also flexible and can be adapted to many different situations.

Note that if you follow this dissemination procedure you will get a 5STARS RD[33].

A much more complete and complex vision of data sharing can be found, for example, in 5.

3.3 The CDUR procedure to evaluate Research Data

Similarly to the RS CDUR evaluation protocol proposed in Section 2.3, the CDUR protocol for RD evaluation that we propose out in here contains four steps to be carried out in the following chronological order: **C**itation, **D**issemination, **U**se and **R**esearch. The RS CDUR evaluation protocol translates to the RD evaluation context in a straightforward way:

(C) Citation. This step measures to what extent the evaluated RD is well identified as a research output. It is also the step where RD producers are correctly identified as well.

As seen in the dissemination procedure (Section 3.2), a reference to cite the work should be well established. If required in a evaluation process, a complete set of RD metadata (including PIDs) should be provided.

(D) Dissemination. This step measures the quality of the RD dissemination plan, as detailed in the previous Section 3.2.

---

[26]https://creativecommons.org/choose/
[27]https://opendatacommons.org/licenses/
[28]https://spdx.org/licenses/
[29]http://www.gnu.org/copyleft/fdl.html
[30]https://creativecommons.org/choose/
[31]http://artlibre.org/licence/lal/en/LAL
[32]http://en.wikipedia.org/wiki/Persistent_identifier
[33]https://5stardata.info/en/

This is also the step dealing with legal (and ethical) issues 15, 34, 38, 39 related to the producers and rightholders (as established in the **C**itation step) deciding and installing the license(s) for the RD dissemination. It can also take into consideration further legal issues related to the objects under study represented in the RD and their legal contexts (13, section 3).

This is also the step concerning Open Science, as the RD license expresses its sharing conditions; and the step where policy makers should establish the Open Science policies that will be applied in the evaluation process.

Finally, let us recall that the inclusion of the list of related publications, software and data sets and other works mentioned in the dissemination procedure helps to prepare the reproducible science issues that are to be taken into account in the **U**se step.

(U) Use. This step is devoted to the evaluation of the technical data aspects. In particular, this step measures the quality of the RD. The RD usability does not only refer to the quality of the scientific output but also can deal with other matters, such as the provided documentation, tutorials and examples of use for easy and intuitive manipulation, etc.

The relevance of the production scientific context, and of the RD generation process, to maximize reuse and reproducibility has been already emphasized in Reference 13, but indeed, the replicability and reuse steps are highly challenging for RD, even if the RD is correctly disseminated. Evaluation committees should take into consideration the difficulties appearing in this matter.

This is the reproducible science step, where it is measured how the published results obtained with the RD can be replicated and reproduced.

(R) Research. This step measures the impact of the scientific research that has required in an essential way the RD under consideration.

The evaluation of this item should follow whatever standards for scientific research quality in the concerned community.

This is the step where the RD related publications (as described in Section 3.1) come into play, and where the evaluation should consider the difficulty of the addressed scientific problems, the quality of the obtained results, the efficiency of the proposed algorithms and data structures, etc. The RD impact can also be assessed through the research impact of the related publications, and through its inclusion (or use) as a data component in other RD.

To end this section, let us remark that similar considerations for the flexibility of the application of the CDUR RS evaluation protocol do apply for RD. See 13 for a more detailed analysis of the RS CDUR evaluation protocol.

3.4 FAIR Research Data

Remark that, as stated in Section 2.4, FAIR principles have been initially designed for data, so our reflections in that section concerning the connection of the FAIR principles with the CDUR protocol do specially apply here. Indeed, there is a lot of recent work on FAIR data issues, see for example 1, 5, 9, 29, 43 and the references mentioned there. We would like to mention some FAIR assessment tools currently under development, such as the automatic FAIR evaluator (DIGITAL.CSIC) of the EOSC-Synergy project[34] or the data sharing evaluation project[35].

The Table 1 below illustrates some connections between our CDUR RD evaluation proposal and the corresponding FAIR Principles as listed in 1. Let us remark that the correspondences between the CDUR and the FAIR principles are not straightforward. For example, the Citation block is placed at the same level that the F1, F2, F3 principles, which does not mean that such principles include completely all the Citation issues, and conversely, that such F principles deal exclusively with Citation elements. The same can be stated for the Dissemination and the Use CDUR steps. In this table we have chosen those FAIR principles that we consider closer to the corresponding CDUR items.

[34]https://github.com/EOSC-synergy/FAIR_eva
[35]https://hal.archives-ouvertes.fr/hal-01943521

Table 1. This table illustrates the relationships between the FAIR principles and the CDUR RD evaluation protocol proposed in Section 3.3.

CDUR	FAIR Principles 1
(C) Citation The RD is well identified, involving issues concerning: - citation form or reference - metadata (including PIDs)	**To be Findable:** F1. (meta) data are assigned a globally unique and persistent identifier F2. data are described with rich metadata (defined by R1 below) F3. metadata clearly and explicitly include the identifier of the data it describes
(D) Dissemination RD is well disseminated, involving issues concerning: - list of included components - RD licence - RD deposit	**To be Findable:** F4. (meta) data are registered or indexed in a searchable resource **To be Accessible:** A1. (meta) data are retrievable by their identifier using a standardized communications protocol A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure, where necessary A2. metadata are accessible, even when the data are no longer available **To be Interoperable:** I3. (meta) data include qualified references to other (meta)data **To be Reusable**: R1.1. (meta) data are released with a clear and accessible data usage license
(U) Use RD facilitates its reuse, involving: - documentation, tutorials. examples... - reproducibility and replicability issues	To be Interoperable: I1. (meta) data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta) data use vocabularies that follow FAIR principles To be Reusable: R1. meta (data) are richly described with a plurality of accurate and relevant attributes R1.2. (meta) data are associated with detailed provenance R1.3. (meta) data meet domain-relevant community standards
(R) Research Measures the impact of the scientific work	Not applicable

Let us remark that we have not found an equivalent FAIR principle to the Research CDUR as shown in the Table 1 above.

Of course, our perception of the connections enumerated in this table does not mean that we consider both approaches to be equivalent or redundant. Indeed, as remarked in 15, our perspective remains at a conceptual level that, on the one hand does not enter in some details that could be addressed through the FAIR Principles. On the other hand, we think that our foundational, simplified perspective allows to better grasp some of the involved basic problems, helping in this way to imagine a journey towards their solution. Yet, we consider, thanks to the Referees suggestions, that getting deeper into the relationships between the FAIR and CDUR principles could be and interesting and challenging subject for future work.

4. Conclusion

Designing and following best practices for research output dissemination are important steps toward accomplishing the Open Science goals, to render research visible, accessible and reusable 2. We also consider that the current evolution in research evaluation practices will enable the adoption of Open Science methods 13, 44, as well as they will facilitate their integration in every day research activities.

As we have already detailed in our work, RS and RD present many similarities concerning dissemination and evaluation issues. For example, we have included in Section 3.1 a RD definition that has been proposed in 13 and that it is clearly based on a RS definition (see 13, 14 and Section 2.1). Following the same scheme, in Section 3 we have proposed and argued in detail RD dissemination and evaluation procedures grounded in the RS proposed dissemination (Section 2.2 and 10) and evaluation (Section 2.3 and 13) procedures.

It is pending work for the future to analyse the potential extension of this parallelism to other kinds of research outputs that are disseminated under similar conditions as RD and RS, that is, without widely accepted publication procedures involving editors or other external actors and where the dissemination is usually restricted through the hands of the production team (eventually including the selection of platforms or repositories).

Sections 2.4 and 3.4 on FAIR RS and RD develop some reflections on the connections between these principles and the proposed dissemination and evaluation protocols. We consider that our dissemination and evaluation (CDUR) proposals, if followed correctly, may clearly contribute towards a more sound implementation of FAIR principles for RS and RD, as they provide robust instructions for their producers to make them more findable and accessible, as well as arguments to choose suitable dissemination platforms to complete the FAIR framework. Moreover, interoperability and reusability could be also fostered with best documentation practices, such as it is proposed in our dissemination procedure; practices that can be evaluated with our CDUR protocol.

As declared in Section 3.4, we think that it will be very important to devote some future work to further study the similarities and differences, and the mutual benefits, between the FAIR and CDUR approaches.

On another note, we observe that one of the advantages of the CDUR protocols for RS and RD described here is that they separate the evaluation of research aspects from those related to much more technical issues concerning software or data, as these different contexts may involve evaluators with disparate levels of expertise in the corresponding areas. Evaluators can then set priorities and adapt the protocol to the evaluation setting.

Furthermore, we consider that our dissemination and evaluation proposals contribute towards the development of Open Science 2. On the one hand, enhancing open access outputs, as we highlight precisely the steps that deal with licensing issues. On the other hand, because we emphasize the role of best dissemination practices the first two steps of the CDUR protocols, as remarked in Sections 2.3 and 3.3.

As a general reflection related to our study of RD dissemination and evaluation issues, let us remark that the CDUR protocol states that a research output (such as RD or RS) that is to be disseminated, should be identified correctly to increase its visibility, as well as the visibility of its producer team and their research work, in order to make it accessible and reutilizable. We have already highlighted in 13 that one of the roles of the evaluation stages is to improve best dissemination practices, such as best credit, attribution and citation, practices that are still to be widely adopted:

> *… we consider that it is in the interest of the research communities and institutions to adopt clear and transparent procedures for the evaluation of research software. Procedures like the proposed CDUR protocol facilitate RS evaluation and will, as a consequence, improve RS sharing and dissemination, RS citation practices and, thus, RS impact assessment.*

After the study presented in this article, it seems to us clear enough that the same statement also applies for RD.

As a final conclusion of our work, we would like to emphasize the underlying RD dissemination/evaluation loop (see 45): first, the CDUR protocol points out to the research community the need to correctly disseminate outputs, as only well disseminated outputs are potential subject of evaluation; secondly, the CDUR protocol also implies that outputs are to be disseminated following the adopted evaluation policies.

In this imbricated context, it is the intention of this work to contribute towards improving dissemination and evaluation procedures, and thus, to enhance best Open Science every day practices.

Data availability

Underlying data

Data underlying the arguments presented in this article can be found in the references and footnotes.

Acknowledgments

With many thanks to the Referees, to the Departamento de Matemáticas, Estadística y Computación de la Universidad de Cantabria (Spain) for hospitality, and to Prof. T. Margoni for useful comments and references.

References

1. Wilkinson M, Dumontier M, Aalbersberg I, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**: 160018.
PubMed Abstract | Publisher Full Text

2. Gomez-Diaz T, Recio T: **Towards an Open Science definition as a political and legal framework: on the sharing and dissemination of research outputs.** *POLIS N.* 2020; **19**. Last Version dated 28/02/2021 is available on Zenodo.
Publisher Full Text | Publisher Full Text | Reference Source

3. European Commission: **Directorate-General for Research and Innovation, Open Science Skills Working Group Report: Providing researchers with the skills and competencies they need to practise Open Science.** 2017.
Reference Source

4. Task Group on Data Citation Standards and Practices, C.-I: **Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data.** *Data Sci J.* 2013; **12**: CIDCR1–CIDCR7.
Publisher Full Text

5. European Commission: **Directorate-General for Research and Innovation. European Commission Expert Group on FAIR Data (2018) Turning FAIR into reality. Final Report and Action Plan from the European Commission Expert Group on FAIR Data.** 2018.
Reference Source

6. Bechhofer S, De Roure D, Gamble M, *et al.*: **Research objects: Towards exchange and reuse of digital knowledge.** *Nature Proceedings 2010.* 2010.
Publisher Full Text

7. Goodman A, Pepe A, Blocker AW, *et al.*: **Ten simple rules for the care and feeding of scientific data.** *PLoS Comput Biol.* 2014; **10**(4): e1003542.
PubMed Abstract | Publisher Full Text

8. European Commission: **Horizon Europe Programme Guide.** 2021.
Reference Source

9. Jacobsen A, de Miranda AR, Juty N, *et al.*: **FAIR Principles: Interpretations and Implementation Considerations.** *Data Intell.* 2020; **2**(1-2): 10–29.
Publisher Full Text

10. Gomez-Diaz T: **Free software, Open source software, licenses. A short presentation including a procedure for research software and data dissemination.** 2014. Presented at the Workshop on open licenses: Data licencing and policies, EGI Conference 2015, Lisbon, May 2015. Spanish version: Software libre, software de código abierto, licencias. Donde se propone un procedimiento de distribución de software y datos de investigación, Septembre 2015.
Reference Source | Reference Source | Reference Source

11. Gomez-Diaz T: **Article vs. Logiciel: questions juridiques et de politique scientifique dans la production de logiciels. 1024 - Bulletin de la société informatique de France.** 2015; 5. First version initially published in the platform of the PLUME project, October 2011.
Publisher Full Text | Reference Source | Reference Source

12. Gomez-Diaz T, Romier G: **Research Software management Plan Template V3.2. Projet PRESOFT, Bilingual document (FR/EN).** 2018.
Reference Source

13. Gomez-Diaz T, Recio T: **On the evaluation of research software: the CDUR procedure [version 2; peer review: 2 approved].** *F1000Res.* 2019; **8**: 1353. First published: 05 Aug 2019.
PubMed Abstract | Publisher Full Text

14. Gomez-Diaz T, Recio T: **Open comments on the Task Force SIRS report: Scholarly Infrastructures for Research Software (EOSC Executive Board, EOSCArchitecture).** *Research Ideas and Outcomes.* **7**: e63872.
Publisher Full Text

15. Gomez-Diaz T, Recio T: **Research Software vs. Research Data I: Towards a Research Data definition in the Open Science context.** 2021. F1000Res.
Reference Source

16. Borgman CL: **The conundrum of sharing research data.** *J Am Soc Inf Sci Tec.* 2012. **63**: 1059–1078.
Publisher Full Text

17. Druskat S, Bast R, Hong NC, *et al.*: **A standard format for CITATION files.** The Software Sustainability Institute. 2017.
Reference Source | Reference Source

18. Smith AM, Katz DS, Niemeyer KE, *et al.*: **Software citation principles.** *Peer J Comput Sci.* 2016; **2**: e86.
Publisher Full Text

19. Fogel K: **Producing Open Source Software: How to Run a Successful Free Software Project.** 2005.
Reference Source

20. Hasselbring W, Carr L, Hettrick S, *et al.*: **From FAIR research data toward FAIR and open research software.** *Information Technology.* 2020; **62**(1): 39–47.
Publisher Full Text

21. Gomez-Diaz T: **Le Projet PLUME et le paysage actuel des logiciels de la recherche dans la science ouverte.** *Zenodo preprint.* 2019.
Publisher Full Text

22. Aimé T: **A Practical Guide to Using Free Software in the Public Sector.**
Reference Source

23. Monteil A, Gruenpeter M, Sadowska J, *et al.*: **« « Garantir la cohérence des données constitue le cœur de notre activité » : entretien autour des enjeux descriptifs du code source».** *Bulletin des bibliothèques de France (BBF), 2021-1.* En ligne.
Reference Source | Reference Source

24. Mons B, *et al.*: *Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud.* 1 Jan. 2017: 49–56.
Reference Source

25. Lamprecht AL, Garcia L, Kuzak M, *et al.*: **Towards FAIR Principles For Research Software.** *Data Science.* 2020; **3**(1): 37–59.
Publisher Full Text

26. Katz DS, Gruenpeter M, Honeyman T: **Taking a fresh look at FAIR for research software.** *Patterns.* March 12, 2021; **2**(3): 100222.
PubMed Abstract | Publisher Full Text | Reference Source

27. Hong C, Neil P, Katz DS, *et al.*: FAIR Principles for Research Software (FAIR4RS Principles) (1.0). 2022.
Publisher Full Text

28. Research Data Alliance/FORCE11 Software Source Code Identification WG, Allen A, Bandrowski A, *et al.*: *Use cases and identifier schemes forpersistent software source code identification (V1.1).* Research Data Alliance. 2020.
Publisher Full Text

29. Ansone SA, McQuilton P, Rocca-Serra P, *et al.*: **FAIRsharing as a community approach to standards, repositories and policies.** *Nature Biotechnology.* 2019; **37**: 358–367.
PubMed Abstract | Publisher Full Text

30. European Parliament and the Council: **Directive 96/9/EC of 11 March 1996 on the 1248 legal protection of databases.**
Reference Source

31. Journal officiel de la République française: **Lois et décrets: Code de la propriété intellectuelle, Version en vigueur au 23 juin 2021.**
Reference Source

32. Guibault L, Wiebe A: **Safe to Be Open: Study on the Protection of Research Data and Recommendations for Access and Usage.** Universitätsverlag Göttingen; 2014.
Publisher Full Text | Reference Source

33. de Cock BM, van Dinther B, Jeppersende Boer CG, *et al.*: **The Legal Status of Research Data in the Knowledge Exchange Partner Countries.** *Knowledge Exchange Report.* 2011.
Reference Source

34. Labastida I, Margoni T: **Licensing FAIR Data for Reuse.** *Data Intelligence.* 2020; **2**(1-2): 199–207.
Publisher Full Text

35. European Commission: **Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information C/2018/2375.**
Reference Source

36. Science Europe: **Presenting a framework for discipline-specific research data management.** *Science Europe Guidance Document D/2018/13.324/1.* 2018.
Reference Source

37. Maurel L: **La réutilisation des données de la recherche aprés la loi pour une République numérique. La diffusion numérique des données en SHS - Guide de bonnes pratiques éthiques et juridiques.** *Presses Universitaires de Provence.* 2018.
Reference Source

38. Boistel R, Bordignon F, Maurel F: **Aspects juridiques de la gestion et du partage Des données.** *Journées Nationales de la Science Ouverte.* 2019. Paris, France.
Reference Source

39. Stérin AL: **Diffuser des données de la recherche dans le respect du droit et de l'éthique: Comment faire lorsqu'on n'est pas juriste ? Guide de bonnes pratiques éthiques et juridiques.** *Presses Universitaires de Provence.* 2018.
Reference Source

40. Callaghan S: **Preserving the integrity of the scientific record: data citation and linking.** *Learned Publishing.* 2014; **27**: S15–S24.
Publisher Full Text

41. Data Citation Synthesis Group: **Joint Declaration of Data Citation Principles.** Martone M, editor. San Diego CA: FORCE11; 2014.
Reference Source

42. DaTaCite Metadata Working Group: **DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.3.** *DataCite e.V.* 2019.
Reference Source

43. FAIR Data Maturity Model Working Group: **FAIR Data Maturity Model. Specification and Guidelines (1.0).** 2020.
Reference Source

44. Guédon JC, Jubb M, Kramer B, *et al.*: **Future of Scholarly Publishing and Scholarly Communication.** *Report of the Expert Group to the European Commission.* 2019.
Publisher Full Text

45. Gomez-Diaz T, Recio T: **Research Software and Research Data: dissemination, evaluation and reusability in the Open Science context.** *17th International Digital Curation Conference (IDCC22). Zenodo.* 2022.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 28 November 2022

https://doi.org/10.5256/f1000research.138948.r152672

✔ **Charles Romain**
Department of Chemistry, Molecular Sciences Research Hub, Imperial College London, London, UK
Henry S. Rzepa
Department of Chemistry, Molecular Sciences Research Hub, Imperial College London, London, UK

The authors addressed the various comments and recommendations by adding text or table. We recommend for indexing.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: FAIR data, workflows for data publication, chemistry

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 07 October 2022

https://doi.org/10.5256/f1000research.138948.r152673

✔ **Mark Leggott**
Digital Research Alliance of Canada, Ottawa, ON, Canada

I feel the authors have addressed my comments, either via edits, or explaining why they feel the specific comment does not apply (e.g. SBOMs).

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Research data management, research software, digital research infrastructure.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 26 April 2022

https://doi.org/10.5256/f1000research.82451.r125654

?

Mark Leggott
Digital Research Alliance of Canada, Ottawa, ON, Canada

In general I found the intent of the article (to propose a common rubric for making data and SW adhere to Open Science and FAIR Principles) to be a reasonable goal, but I'm not sure the article clearly achieves that goal. The authors state that "Bearing in mind the above described landscape, the goal of our work here is to contribute to the improvement of the scientific endeavor with protocols that could help researchers, and the community at large, in the dissemination of their produced RD and RS, while contributing to the accomplishment of Open Science goals." I don't feel that the proposed protocols (CDUR) provide sufficiently detailed recommendations to support this goal.

Part of the challenge I have in saying this is that work of initiatives like FAIR4RS would be very informative in this article, but they are not mentioned. Also, the FAIR4RS Principles have direct intersections with some of the CDUR approaches, which seems overly simplified in view of this FAIR4RS rubric. I find it unusual that the paper does not even mention the work of the FAIR4RS Working Group, which has articulated a number of the concepts and approaches highlighted in this article. The work of the RDA-FORCE11 Software Source Code Identification Working Group is also of direct relevance to the PID discussion in 2.4, so should be highlighted.

A table comparing the CDUR recommendations against data and SW might be useful, as it would extract the key elements of the proposed approach and make it easier for the reader to make the connections.

The authors could mention the value of a Software Bill of Materials (SBOMs) in section 2.1, para 4 (Moreover...) Mentioning how a SMP can be integrated with a DMP, reinforcing the idea that the data and SW can be considered in a common rubric, would also be beneficial. There has been some work by the Software Sustainability Institute to develop a SMP meant to be integrated with standard DMP tools.

I find the idea of drawing connections between the practices needed to support similar Open Science/FAIR concepts with data and SW is very desirable, and the authors do provide one of the few attempts to articulate this. If they were able to achieve a better integration of additional and specific resources and best practices with their CDUR approach, it would benefit the article substantially.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Research data management, research software, digital research infrastructure.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 08 Sep 2022

Teresa Gomez-Diaz, CNRS, Paris-Est, France

Dear Mark Leggott,

Many thanks for all your comments that have helped us a lot to improve our work. Here are some answers while we are preparing a new version.

- [intent of the article]
- [proposed protocols (CDUR) provide sufficiently detailed recommendations]
Of course, it is our opinion that the article achieves that goal to a sufficient extent as to merit its approval by the Referees.

- [initiatives like FAIR4RS...]
Perhaps you have not noticed that in the original version of this paper there were specific references to FAIR RS in References 19,20,21. In the new version we have added more recent ones following your suggestion. We will also include Research Data Alliance/FORCE11 Software Source Code Identification WG, Allen, Alice, Bandrowski, Anita, et al. (2020). Software Source Code Identification Use cases and identifier schemes for persistent software source code identification (1.1). https://doi.org/10.15497/RDA00053. Many thanks for this suggestion.

- [table table comparing the CDUR recommendations against data and SW]
Thanks for your suggestion, but we do not think a comparison table is necessary, as CDUR for RS and for RD find pretty similar formulation, only a new item for RD dissemination (item 3 in Section 3.2) has been added to the RS dissemination protocol to highlight the potential difficulties concerning RD legal (and ethical) issues.

- [Software Bill of Materials (SBOMs)]
Many thanks for your suggestion, but we consider that software security issues, the main topic of, for example, the report https://linuxfoundation.org/wp-content/uploads/LFResearch_SBOM_Report_020422.pdf, are not part of our conceptual approach of Section 2.1. As already mentioned to other Referees, we do not enter in several concrete and technical issues. The point of software security issues, as well as many other ones, could be considered by the evaluation committees in the Use step of the CDUR protocol, which can be easily adapted by the committees to consider such technical points. Indeed, there are many relevant issues that are not detailed in our definition and proposed protocols, as we aim to address the subject from a more conceptual perspective. This point of view has been clarified in the new version of [Reference 13]. The decision of considering such broad point of view has required us to include in the new version of [Reference 13] comments reacting to some Referee questions that asked us, like you do, the consideration of different specific issues that they considered we had forgotten.

- [how a SMP can be integrated with a DMP]
Thanks again for your suggestion. This could be the subject of future extension of our work, but it is out of the scope of the present article.

- [I find the idea of drawing connections between the practices needed to support similar Open Science/FAIR concepts with data and SW is very desirable, and the authors do provide one of the few attempts to articulate this.]
Thank you so much for this very positive comment, that in some sense supports our perception that this article already provides enough contributions to merit its publication as it is, although, there is always room for improvements in future work.

- [achieve a better integration of additional and specific resources and best practices]
Following your suggestion we provide a new version of sections 2.4 and 3.4 regarding additional reflections on the relationships between FAIR and CDUR issues. For more detailed information on the CDUR protocol the appropriate reference is [Reference 11], while in the present paper we have included only a short description to explain that CDUR

can be applied in a similar way for RS and for RD.

Teresa Gomez-Diaz, Tomas Recio

Competing Interests: No competing interests were disclosed.

Reviewer Report 07 February 2022

https://doi.org/10.5256/f1000research.82451.r121544

?

Charles Romain
Department of Chemistry, Molecular Sciences Research Hub, Imperial College London, London, UK
Henry S. Rzepa
Department of Chemistry, Molecular Sciences Research Hub, Imperial College London, London, UK

This manuscript aims to highlight similarities between research software (RS) and research data (RD) in the context of Open Science. The authors propose protocols and procedures for the dissemination and its evaluation for both RS and RD. The introduction provides a clear overview of the context by citing relevant sources. The manuscript then discusses definition, dissemination procedures and evaluation protocols for both RS and RD.

The comparison of RS and RD in the context of dissemination (and its evaluation) is interesting and highlights some similarities in the bottlenecks and challenges that need to be addressed.

However, our main general concern is the high heterogeneity of RD compared to RS (maybe a naïve vision) which makes these general procedures and protocols not always applicable or too vague to be practically useful as they don't provide practical solutions.

However, they have the merit to draw the attention to the importance of following good practices (e.g. protocols, procedures) to disseminate RD and RS, and provide general guidance which can help to identify what could be some potential practical solutions. We recommend this article to be indexed as a good basis for further discussions.

Below we provide some detailed comments and suggestions on specific points discussed in the manuscript that could be included in future version of the article.

Proposed definitions:

- The proposed definitions used in this article are based on another article currently awaiting peer review[1]. The authors should update these definitions after comments from the reviewers and approval of the other manuscript.

- We generally agree with the proposed definitions. However, regarding the definition of RD we would add that following protocols and standards established in the field are important. For example, "*the data has been collected, processed and analysed* ***following protocols, procedures and standards*** *established in the field to produce a results…..*"

Dissemination and evaluation procedures:

- The dissemination procedure and the CDUR protocol for RS have previously been reported and reviewed, we don't have further comments on these.

- One important aspect and challenge for RD dissemination that should be mentioned is to make data "machine-readable" to facilitate collection and enable re-use, and thus contribute to new data-driven discoveries (e.g. using machine-based tool). In some fields, RD are mainly shared as "human-readable" format only, usually in a monolithic supporting information document along with a scientific publication (e.g. PDF format). RS in contrast has well developed repositories which to a certain extent are machine actionable, as well as being closely integrated into the publication processes (such as Github via https://github.com/openjournals and https://joss.theoj.org). Overall, we think this article should have a paragraph where the state of play might be on machine-readable or even machine-actionable RD and RS and perhaps comparing how this is evolving for both of them.

- In general, it would be good to further emphasise that many aspects discussed in these procedures and protocols can easily be addressed with relevant metadata which are keys to address the FAIR principles. In terms of metadata, this plays a key role in ensuring Findability/Discovery of the object, using the metadata registry MDS (metadata store). It would be good for this article to perhaps illustrate the role of registered metadata in finding both RD and RS, and perhaps to explore the granularity of the metadata for both. Is it good enough to use metadata purely to discover the functionality of a RS code, or might it be necessary to explore in more details the functions and libraries in RS?

- It is maybe a naïve vision, but research data features a much broader heterogeneity than RS, so many different types of data can be generated that it makes these procedures difficult to apply or generalise to RD in general (As mentioned in the report of reviewer 1 for RS[2]).

- In general, more specific examples are provided for RS than for RD. More examples or references for RD would be useful for the readers.

Dissemination steps:

- Contrary to RS, versioning is rare or can be difficult to achieve with RD. Complementary data can be generated to support new conclusions but does not necessarily feature new versions of the previous ones, thus "*revisiting version release"* as mentioned doesn't necessarily make sense, but *each time new data are generated* would be more appropriate.

- A single name or title for RD is not necessarily as straightforward as it is with RS, as RD usually are made available along with scientific articles (as mentioned in the three main characteristics), the name of the dataset can be related to the title of the narrative it accompany though.

- "*Choose a web site, forge…*" Data repository should be explicitly mentioned. In addition,

selecting data repository which generate relevant metadata for the discipline (if available) should be considered (rather than generic data repository).

- The address of contact can easily be addressed via PID associated with the authors (i.e. ORCiD).

Evaluation steps:

- We would encourage the author to discuss how the different C, D, U, R steps of the protocol help to meet the FAIR criteria. To some extent, the Citation step help to address the Findability, the Use step deals with the interoperability, etc...

Data citation:

- These can easily be evaluated via the attribution of DOI, and PID for the researchers (OrCiD). The data citation can follow some guideline previously defined by various organisation (see joint declaration for data citation - FORCE11) and should be given as examples.

The choice of file format (e.g. non-proprietary, open format, machine-readable) to enable re-use is important and something that should be evaluated in the protocol.

The replicability and reuse steps are highly challenging for RD, even if RD are “well-disseminated”. We would emphasize the importance of the context and how the RD have been generated, to maximize reuse and reproducibility.

References

1. Gomez-Diaz T, Recio T: Research Software vs. Research Data I: Towards a Research Data definition in the Open Science context. *F1000Research*. 2022; **11**. Publisher Full Text
2. Gomez-Diaz T, Recio T: On the evaluation of research software: the CDUR procedure. *F1000Research*. 2019; **8**. Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

No source data required

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: FAIR data, workflows for data publication, chemistry

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 08 Sep 2022

Teresa Gomez-Diaz, CNRS, Paris-Est, France

Dear Charles Romain and Henry S. Rzepa,

Many thanks for all your comments that have helped us a lot to improve our work. Here are some answers while we are preparing a new version.

- [high heterogeneity of RD compared to RS]
We will include some explanations. Indeed, heterogeneity appears obviously between software and data, yes. But if we keep these aspects aside and focus on the similarities, we can learn a lot from the common features that appear in the production context of RD and RS. This is a general procedure that can be adapted to several situations. Even when the differences are too important, and maybe the proposed dissemination and evaluation procedures are not directly applicable as such in both settings, they could help to suggest hints to address the diverse issues appearing in each environment.

- [to draw the attention to the importance of following good practices]
Many thanks!

- [Proposed definitions, article currently awaiting peer review]
We are preparing Version 2 of [Reference 13] bearing in mind Referees' comments.

- [Proposed definitions, would add following protocols and standards]
We agree with you, but we do not think that is necessary to change the formulation of our definition. Yet, in the new version of [Reference 13], we have detailed and emphasized the relevance, in order to characterize the RD concept, of the research team decisions in the scientific production framework, which includes the *protocols, procedures and standards to be followed* in order to consider some given data as true RD. See, for example, our reference to the STRENDA standards developed for investigations on enzyme activities. Some of your comments to the present paper have helped us to improve the new version of [Reference 13], as they are related to the conceptual framework. For this reason, the new version of the present paper does not reflect them in here.

- [Dissemination and evaluation procedures, machine-readable...]
Yes, there are many relevant issues that are not detailed in our definition and proposed protocols, as we aim to address the subject from a more conceptual perspective. This point of view has been clarified in the new version of [Reference 13]. The decision of considering such broad point of view has required us to include in the new version of [Reference 13] comments reacting to some Referee questions that asked us, like you do, the consideration of different specific issues that they considered we had forgotten.

- [Dissemination and evaluation procedures, a paragraph where the state of play might be on machine-readable or even machine-actionable RD and RS...]
Again, out of our scope/goals, we agree that machine actionable is an important issue, but we do not enter in this kind of very concrete aspects.
Some of your comments to the present paper have helped us to improve the new version of [Reference 13], as they are related to the conceptual framework. For this reason, the new version of the present paper does not reflect them in here.

- [Dissemination and evaluation procedures, FAIR principles....]
We deal with FAIR issues Sections 2.4 and 3.4, and as remarked there, we do remain in a conceptual level that does not enter in many challenging issues like the ones you mention.

- [Dissemination and evaluation procedures, research data features a much broader heterogeneity than RS]
See our comment above about heterogeneity issues.

- [Dissemination and evaluation procedures, more specific examples are provided for RS than for RD]
It seems to us that the targeted research audience is used to handle RS and/or RD as part of their everyday research practices, so they do not require further examples. Anyway, we have made references to different aspects of linguistic, environmental and geographical data in Section 3 of [Reference 13]. Moreover, further examples can be found easily in the literature, as can be seen in the bibliography included at the end of both works (the present article and [Reference 13]) .

- [Dissemination steps, RD versioning is rare]
References 49 and 50 of [Reference 13] do mention data versioning as the standard term. Indeed there are cases where versioning could be a difficult issue. But we consider that the RD production team can select the data that is to be released, and maybe provide several versions of the same RD object, thus, “version” seems to be the correct term. The inclusion of the sentence “each time new data are generated” could generate confusion: are we are referring to the same RD object or not? This is similar to what happens to new versions or main releases of software.

- [Dissemination steps, single name or title for RD is not necessarily as straightforward]
When we speak about the publications associated to RS or RD this means the place where the obtained results are published. Several articles or other scientific contributions (e.g. conferences, books...) can present the scientific results obtained with the RS and/or the RD. These objects (articles, RS, RD) can have different names. See for example the list of publications related to SageMath-Combinat [https://www.sagemath.org/library-publications-combinat.html] or the RD entitled “Vital Statistics data” associated to the publication [Recio Alcaide A, Pérez López C, Bolúmar F. Influence of sociodemographic factors in birth seasonality in Spain. Am J Hum Biol. 2022 Aug 8:e23788. doi: 10.1002/ajhb.23788].
It is usually difficult to foresee if some lines of code, or some collected data will become few months later more organized RS or RD. The fact of giving a name to this initial object is the

first step to identify it as a research output.

- [Dissemination steps, Data repository should be explicitly mentioned]
Yes, you are right. This point has been studied in the Conclusions of [Reference 13], that is the Where? question of the RD Borgman's conundrum challenges. We will modify in the new version of the present article the RD dissemination procedure to include your suggestion.

- [Dissemination steps, The address of contact can easily be addressed via PID associated with the authors (i.e. ORCiD).]
Well, we do not fully agree with this optimistic view, as ORCID usually refers to one researcher, while a mail address can refer to a (evolving) team, which is a much more generic solution. In our opinion, the identification issue can still be a quite challenging issue, mainly for software or data that have been developed for many years collectively. Thus, we do prefer to keep the current formulation in Sections 2.2 and 3.2 of the present article.

- [Evaluation steps, how the different C, D, U, R steps of the protocol help to meet the FAIR criteria]
Yes, you are right, and many thanks for this interesting suggestion, that have addressed in sections 2.4 and 3.4 of the new version.

- [Data citation, can easily be evaluated via the attribution of DOI..]
Citation examples are already available in the cited works, see for example References 4,34,35,37, and we will add this one: Altman M, Crosas M. The evolution of data citation: From principles to implementation . IASSIST Quarterly. 2013;37]. See also our comments above concerning ORCID issues.

- [choice of file format]
Yes, but we think that this issue is already addressed – perhaps not in such level of detail as you mention, as this level does not correspond to the more conceptual approach of our work – , as indicated in the dissemination protocols, in Sections 2.2 and 2.3:
Good metadata and respect of open standards are always important when giving away new components to a large community: it helps others to reuse your work and increases its longevity.
And this is to be evaluated in the CDUR protocol in the Dissemination step, as the use of open formats corresponds to Open Science issues. Other technical issues can be evaluated in the Use step, see the description of this step in section 4.2 of [Reference 11].

- [The replicability and reuse steps are highly challenging for RD, even if RD are "well-disseminated". We would emphasize the importance of the context and how the RD have been generated, to maximize reuse and reproducibility.]
Yes, see [Reference 13] where the importance of the context, that is intimately related to the data and RD concepts, is examined thoroughly in Section 3 (OECD Glossary of Statistical Terms and Reference 18 indicated in [Reference 13]. Following your suggestion, this point has been highlighted now in the new version in preparation, see the CDUR protocols proposed for RD (Section 3.3).

Teresa Gomez-Diaz, Tomas Recio

Competing Interests: No competing interests were disclosed.