Podcast Code for Thought: Research Software and Research Data in Open Science

Teresa Gomez-Diaz, Tomas Recio

Podcast Code for Thought*:

Research Software and Research Data in Open Science

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

[1]University Gustave Eiffel-LIGM-CNRS, Est of Paris
[2]University Antonio de Nebrija, Madrid
Contact: Teresa.Gomez-Diaz@univ-eiffel.fr, trecio@nebrija.es

July 18th 2023

Abstract

The goal of this text is to write down and extend the questions and answers that have been prepared for the Podcast Code for Thought registered in May 2023 at the Laboratoire d'informatique Gaspard-Monge (LIGM), located at the Gustave Eiffel University, at the Est of Paris. This podcast goal was to have an informal conversation about the recent published work on Research Software and Research Data in the Open Science context by T. Gomez-Diaz and T. Recio. The interview was conducted by Peter Schmidt, who provided the questions we do answer here, and that were answered by T. Gomez-Diaz in the recorded podcast. The podcast, entitled *Open Data, Open Software - with Teresa Gomez-Diaz*, is now published [1].

Keywords. Research Software, Research Data, Open Science, Research Evaluation, FAIR

1 Foreword

Code for Thought [2] is a podcast series launched by Peter Schmidt [3] in January 2021 [4]. The goals of Code for Thought podcasts are to support, explore and promote the growing community of Research Software Engineers (RSE) in the UK and abroad, and it is realized in close collaboration with the Software Sustainability Institute [5] and the Society for Research Software Engineering [6] as well as with other RSE communities outside the UK.

*Permision is granted by Peter Schmidt for the inclusion of the Code for Tought logo in this document.

1. https://codeforthought.buzzsprout.com/1326658/13216530-en-open-data-open-software-with-teresa-gomez-diaz
2. https://codeforthought.buzzsprout.com
3. https://www.software.ac.uk/about/fellows/peter-schmidt
4. The first Code for Thought podcast *And so it begins...* is available at https://codeforthought.buzzsprout.com/1326658/7121884-and-so-it-begins.
5. https://www.software.ac.uk/
6. https://society-rse.org/

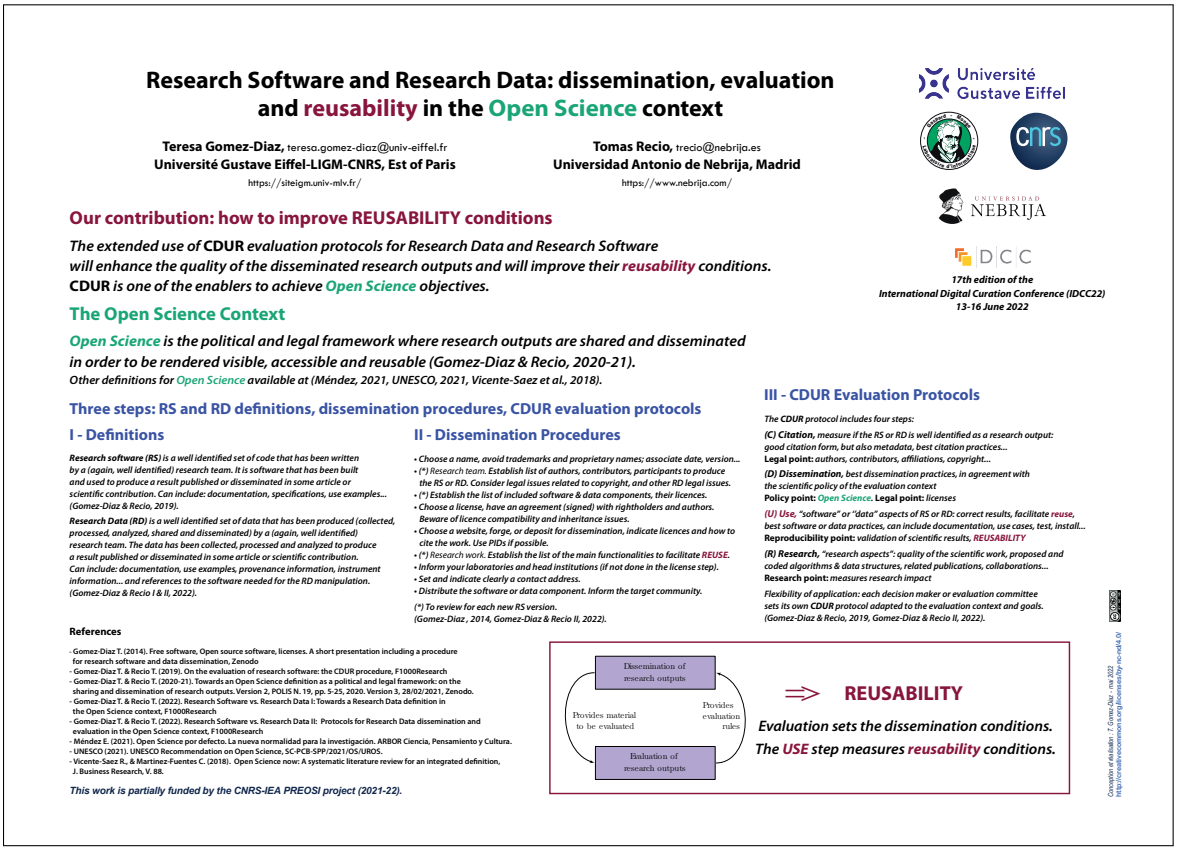

FIGURE 1 – Research Software and Research Data: dissemination, evaluation and reusability in the Open Science context, IDCC22. Work licensed under the CC-BY-NC-ND Creative Commons License.

This particular podcast is dedicated to recent work by T. Gomez-Diaz and T. Recio on Research Software and Research Data in the Open Science context that was published in F1000Research [7] last year [8, 9].

The poster [10] entitled *Research Software and Research Data: dissemination, evaluation and reusability in the Open Science context* summarizes these two articles and further enhances the importance of the work presented there in the context of reusability of these research outputs. It was presented at the 17th edition of the International Digital Curation Conference (IDCC22) that took place in June 2022 [8]. This poster is included in Figure 1.

The questions to prepare the podcast were sent by Peter Schmidt one week before the recording was done and the answers were started to be arranged in a draft document which became a first version of the present document.

The above mentioned publications on Research Data [8, 9, 10] do follow our work on Research Software [5] and on Open Science [6, 7] that was also the object of another Code for Thought podcast: *Open Science and Research Software* [9]. It was also recorded in May 2022 at the Laboratoire d'informatique Gaspard-Monge and released in July 2022. Furthermore, the poster [7] entitled *The future of Open Science asks for a common understanding* was presented at the EGI Virtual Conference 2021 [10] and it is included in Figure 2.

In order to complete the Open Science framework in which we place our work, we would like to mention here the UNESCO Recommendation on Open Science adopted in November 2021 [12].

7. https://f1000research.com/
8. https://www.dcc.ac.uk/events/idcc22
9. https://codeforthought.buzzsprout.com/1326658/10822132-open-science-and-research-software
10. https://indico.egi.eu/event/5464/

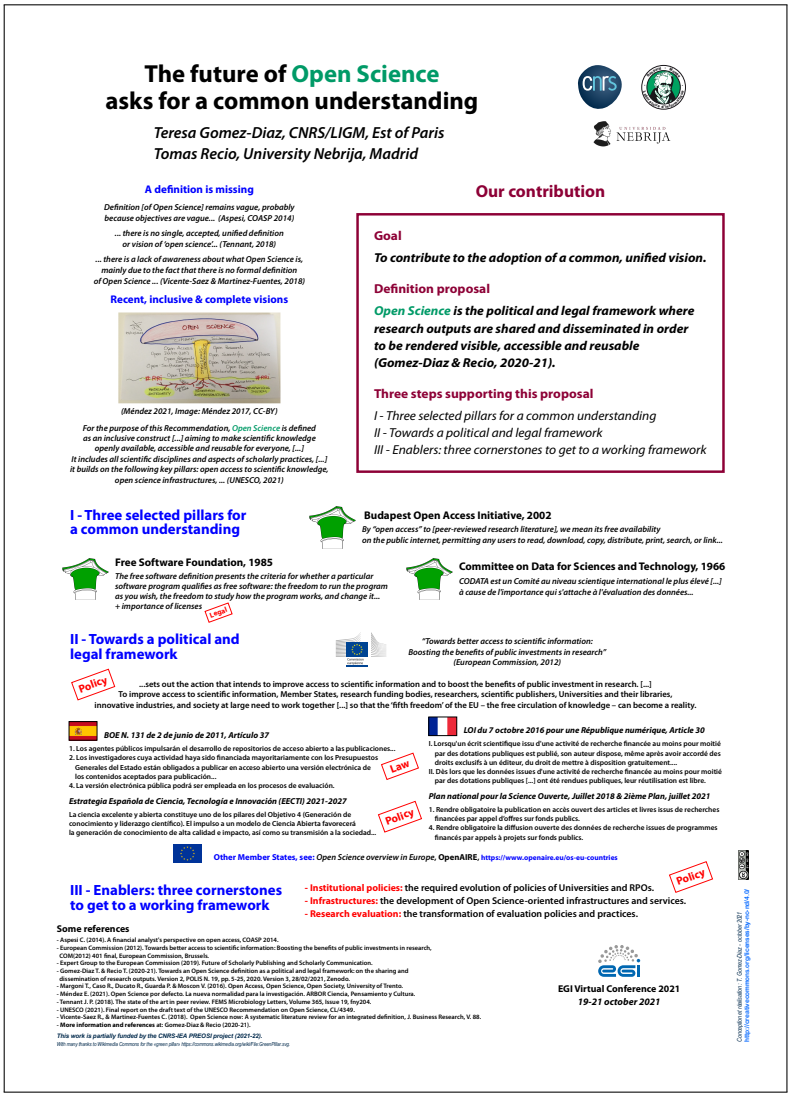

FIGURE 2 – The future of Open Science asks for a common understanding, EGI Virtual Conference 2021. Work licensed under the CC-BY-NC-ND Creative Commons License.

2 Questions and Answers

Q1. The introduction to your paper mentions Borgman's Conundrum. What is the Conundrum, and how does it relate to your paper?

A1. Christine Borgman explained in her article entitled *The conundrum of sharing research data* [1] that to share research data is a difficult problem and asked some fundamental questions: *which Research Data might be shared, by whom, with whom, under what conditions, why, and to what effects.*

In the recently published article [8] we have proposed a *Research Data (RD)* definition and a framework in which we can provide answers to the Borgman's Conundrum questions. The formulation of this RD definition finds a similar formulation that the one also proposed for *Research Software (RS)* in [5], in 2019.

Q2. You are linking Research Software and Research Data. Could you explain the link - i.e. how they are connected.

A2. One link is that both productions, RS and RD, find similar problems in the scientific context. For example there is not an accepted way to publish them, that is, a publication system that is so generally and widely accepted as the one that exists for articles and scientific publications since longtime.

Another important point is that researchers are evaluated only and mainly regarding their publications but still without taking into consideration the RD or RS production. This is changing a lot nowadays, but nevertheless it will take time to arrive to a system so widely adopted and accepted as the one existing for articles.

What we show in these two articles [8, 9] and in the poster [7] (see Figure 1) is that the framework in which we define RS and propose the CDUR evaluation protocols to give value to this production [5], this framework is also valid for RD. Which means that we can propose a similar formulation for the RD and the RS definitions, and similar dissemination [4, 9] and CDUR evaluation protocols [5, 9].

Q3. What are the difficulties to define Research Data

A3. Well, the problem is not to define RD, but to define what data is. Data is a complex object from many points of view: what is data, what is information, what is evidence... It is also complex from the scientific point of view, as data manipulated by a chemist can have little to do with data manipulated by a sociologist.

But the difficulties are also from the legal point of view. One of the points that have helped us is that we do not need to define software. In our RS work [5], our RS definition stands over a software definition that is provided by the legal framework of author rights and copyright (as defined in France, Spain and by the EC). But it has taken us a lot of time to understand that we have the same schema for RD although for different reasons. We do not try to define data, but its legal context is not the same as in software. Software legal context is more or less easy and regards mainly copyright and licenses, and data is far too much complex. In fact legal context of data to study temperature in a region for a year usually does not raise many legal issues. But data can have associated author rights, or maybe not, and there are also other legal issues like personal data, medical data, *sui generis* database rights...

Yet, we can provide a RD definition with similar formulation than the RS definition [8] if we forget about how to define data. Sometimes you need mathematicians to understand and deal correctly with what means a definition and which is the right way to propose one.

Q4. What are the problems you see with current definitions of Research Data. For instance the definition by UKRI

A4. One of the problems is, as far as we understand, that these definitions try to say what data is. But data is a too much complex object, that *should be studied in a case-by-case basis* [3], and this is not only from the legal point of view, but also from the scientific point of view, taking into consideration a specific scientific community, a concrete problem. We have provided examples in [8]: temperature related data, the STRENDA Guidelines for enzyme activities, the linguistic research work at LIGM, what is environment information as defined by the French *Code de l'environnement*.

The Research Data definition coming from the Concordat on Open Research Data that is signed by the research councils of the UK Research and Innovation (UKRI) organisation [2] is the following one:

> ***Research data*** *are the evidence that underpins the answer to the research question, and can be used to validate findings regardless of its form (e.g. print, digital, or physical). These might be quantitative information or qualitative statements collected by researchers in the course of their work by experimentation, observation, modelling, interview or other methods, or information derived from existing evidence. Data may be raw or primary (e.g. direct from measurement or collection) or derived from primary data for subsequent analysis or interpretation (e.g. cleaned up or as an extract from a larger data set), or derived from existing sources here the rights may be held by others.*

As you can see, there is a lot of information in the above text that is related to what data is. A good point of this definition is that the goal is research. But how to provide answers to the Borgman's Conundrum questions? Who shares the data? Which data?...

Q5. What is your definition of Research Data?

A5. The RD definition proposed in [8] is the following:

> ***Research Data*** *is a well identified set of data that has been produced (collected, processed, analyzed, shared & 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 RD 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).*

We can summarize the above definition in the following three main characteristics:
— the goal of the collection and analysis is to do research, that is, to answer a scientific question (which includes the validation of research findings),
— 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.

This definition tackles then with three points: which is the research object (RD), who is the research team and which is the RD related research work to answer a research question. In this context it is easier to provide answers to the Borgman's Conundrum questions, as presented in the Conclusions section of [8].

Q6. Your second paper discusses protocols that can be used to make sure research software and data adhere to FAIR principles. It mentions the CDUR principles. What are they?

A6. The question is how to give value to the RD and/or the RS production.

The CDUR evaluation protocol has been initially formulated for RS in [5] and has been extended to RD in [9]. It has been designed to help evaluators, members of evaluation committees, and evaluated researchers, members of the research teams responsible of the RS or RD.

There are four steps in the proposed evaluation protocol, which is flexible enough to be applied in different evaluation contexts:

(C) Citation. This step measures if the RS or RD are well identified as a research output, i.e. if there is a good citation form, or good metadata. We look here to best citation practices.

This is a legal related point where we ask for authors (if any) are well identified, which are their affiliations, and for example the % of their participation in software writing.

(D) Dissemination. In this point we look to best dissemination practices, in agreement with the scientific policy of the evaluation context. The dissemination of RS and RD needs a license to set the sharing conditions. For RD there are maybe further legal issues to look at (personal data, *sui generis* database rights...).

This is a policy point in which we look at Open Science requirements.

(U) Use. This point examines "software" or "data" aspects, in particular the correct results that have been obtained, and we can also look if their reuse has been facilitated, the output quality, best software/data practices such as documentation, testing, installation or reuse protocols, up to read the code, launch the RS, use examples...

This is the reproducibility point that looks at the validation of scientific the results obtained with the RS and/or the RD.

(R) Research. This point examines the research aspectsassociated to the RS and/or RD production: the quality of the scientific work, the proposed and coded algorithms and data structures, which are the related publications, the collaborations, the funded projects...

This point measures the impact of the RD and/or RS related research.

The CDUR protocol is flexible enough to be adapted to different evaluation situations. Each evaluation committee sets its own protocol adapted to a specific evaluation context with its own goals (recruitment, career evolution, publication...).

It is our understanding that, if followed correctly, CDUR may clearly contribute towards making RS and RD more visible, accessible and reusable, but also towards providing tools to build more solid implementations of FAIR principles for RS and RD [9].

Q7. How do the practical CDUR steps differ for Research Software and Research Data

A7. As you can see in the previous answer, the formulation is pretty similar. For example, in the **(D) Dissemination** step, the licenses that should be used for RS or RD are not the same, but there is the need of a license in both cases to set the conditions in which the outputs are shared and disseminated [4].

The **(U) Use** step is the one in which we concentrate the software or the data specific issues, but it is the same step for both outputs, to deal basically with quality and reuse issues but also with the reproducibility issues of the related scientific work. This step is intimately related to the Reusable FAIR principle [9].

Q8. From your perspective: how big is the problem, that is, how much research software and research data do NOT follow Open Science/FAIR principles

A8. This is still a big problem!! There are still a lot of RS and RD that do not follow Open Science/FAIR principles!! As said simply in a recent OpenAire webinar [11] by Kostas Glinos (April 2023): *you get what you reward*. Or as we have explained in the poster included in Figure 1 [10], there is an *evaluation/dissemination loop*. The dissemination of the research output will follow evaluation rules, and the dissemination provides outputs to be evaluated. If the evaluation rules evolve, the dissemination of the research outputs will adapt to the new rules.

There is now, as far as we understand, a largely accepted position to say that research evaluation should be changed, in particular to include other outputs than publications, namely RS and RD. But how? A lot of work is ongoing now to study how to propose these changes in the Open Science context, for example with the Coalition for Advancing Research Assessment (COARA) [12] work, a recent coalition that gathers many Universities and other institutions.

There is need for a new, large consensus on these changes in order to be adopted extensively. Our work [5, 9] proposes precise RS and RD evaluation protocols which can help to build a new research assessment system and, therefore, they can help to foster Open Science/FAIR policies and best practices.

Q9. Do you still encounter scientists and research software engineers who are NOT familiar with these principles?

A9. Truly, the construction, development and implementation of these principles are coming in the context of very recent, ongoing work, and changes take time in the scientific community. Some research communities, institutions, countries can go faster than others to adopt and install these changes. But it is clear that there is need for a large, general consensus, and the COARA coalition can provide the necessary context for this.

Q10. Are there people opposed to the principles?

A10. As far as we understand, most of the people are not opposed to Open Science or FAIR principles, many do not know them very well, or if you know them, how to comply? And how much time should you spend for this? And which are the consequences for the career (evaluation...)

11. https://www.openaire.eu/blogs/openaire-in-greece-and-cyprus-webinar-on-the-european-reform-on-research-assessment
12. https://coara.eu/

Q11. France has legislation to make research software (and data?) for publicly funded projects Open Source. Do you think this legislation has helped? Is it enough?

A11. Legislation for data is in the Digital Republic Law [13] (October 2016), that gives a new duty for Universities and other research performing organizations. Its Article 6 says that *open data should be the default for all publicly funded data, including research.* So, for example, if your research project is publicly funded for at least half of the total funding, the data produced within the project should be open.

For software it was a bit more unclear, but the Second National Plan for Open Science of the Minister of Higher Education, Research and Innovation includes RS (July 2021) [11].

This new legal and political context helps to change the system, but still the system needs deep changes like, for example, in the evaluation, to consider these productions and how well they are disseminated to facilitate their reuse and the validation of the associated research findings.

Q12. What do you and your colleagues do to promote CDUR/FAIR further

A12. At the moment we try to make this work [5, 6, 7, 8, 9, 10] better known, presenting it to several conferences etc. This is why we thank you so much, Peter, as well as the Code for Thought team, for helping us to promote our work and to give us this opportunity to talk about it.

There is also a very recent news in the F1000Research journal, as there is a new collection of publications called *Innovations in Research Assessment Collection* [14] and our three articles [5, 8, 9] have been included there last week. This will help a lot to give more visibility to this recent work, and, again, we thank the Collection Advisors, Simon Hettrick and David Moher, for the inclusion of our articles in this collection.

3 Epilog

This document reviews the questions and answers that were the basis of the podcast Code for Thought on Research Software and Research Data in Open Science and that was recorded in May 2023 at the Gustave Eiffel University. The podcast goal was to have an informal conversation about our recent published work and the interview was conducted by Peter Schmidt, who provided the questions. Our intention here is not to realize an accurate transcription of this podcast, but rather to complete and accompany the recorded audio, and to extend and clarify few points that could be difficult to grasp or that we forgot to mention there.

The work presented in the podcast and in this document has been partially funded, among others, by the LIGM, the University Gustave Eiffel, and the CNRS-International Emerging Action (IEA) PREOSI (2021-22). It has also enjoyed from the hospitality of the Departamento de Matemáticas, Estadística y Computación de la Universidad de Cantabria (Spain).

This work was also presented at the *Journée de clôture du printemps de la donnée*, June 30th 2023 [15].

We hope that you have enjoyed listening to the podcast and reading this text, many thanks for listening and/or reading about our work!

References

[1] Borgman CL: The conundrum of sharing research data. J. Am. Soc. Inf. Sci. Technol. 2012; 63: 1059-1078. https://doi.org/10.1002/asi.22634

[2] Concordat on Open Research Data, that was signed by the research councils of the UK Research and Innovation (UKRI) organisation, 2020, https://www.ukri.org/wp-content/uploads/2020/10/UKRI-020920-ConcordatonOpenResearchData.pdf

[3] 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. https://repository.jisc.ac.uk/6280/

13. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000033202746/
14. https://f1000research.com/collections/innovations-in-research-assessment/about-this-collection
15. https://printempsdeladonnee.fr/events/journee-de-cloture-du-printemps-de-la-donnee/

[4] Teresa Gomez-Diaz. 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, https://zenodo.org/record/11709
Spanish version: Software libre, software de código abierto, licencias. Donde se propone un procedimiento de distribución de software y datos de investigación, Septembre 2015, https://zenodo.org/record/31547

[5] Teresa Gomez-Diaz, Tomas Recio. On the evaluation of research software: the CDUR procedure, F1000Research 2019, 8:1353, https://doi.org/10.12688/f1000research.19994.2

[6] Teresa Gomez-Diaz, Tomas Recio. Towards an Open Science definition as a political and legal framework: on the sharing and dissemination of research outputs.
V3 dated 28th February 2021 is available at https://zenodo.org/record/4577066
V2 dated December 2020 is published in POLIS N. 19, https://uet.edu.al/polis/wp-content/uploads/2022/01/polis-19.pdf
V1 dated September 2020 and entitled *A policy and legal Open Science framework: a proposal* is available at https://zenodo.org/record/4075106

[7] Teresa Gomez-Diaz, Tomas Recio. Poster. The future of Open Science asks for a common understanding, EGI Virtual Conference 2021, https://zenodo.org/record/6433533

[8] Teresa Gomez-Diaz, Tomas Recio. Research Software vs. Research Data I: Towards a Research Data definition in the Open Science context, F1000Research 2022, 11:118, https://doi.org/10.12688/f1000research.78195.2

[9] Teresa Gomez-Diaz, Tomas Recio. Research Software vs. Research Data II: Protocols for Research Data dissemination and evaluation in the Open Science context, F1000Research 2022, 11:117, https://doi.org/10.12688/f1000research.78459.2

[10] Teresa Gomez-Diaz, Tomas Recio. Poster. Research Software and Research Data: dissemination, evaluation and reusability in the Open Science context, 17th International Digital Curation Conference 2022 (IDCC22), https://zenodo.org/record/6778872

[11] Minister of Higher Education, Research and Innovation, Second National Plan for Open Science of the Research Ministry (July 2021),
https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-2021-2024-vers-une-generalisation-de-la-science-ouverte-en-48525

[12] UNESCO Recommendation on Open Science, November 2021, https://unesdoc.unesco.org/ark:/48223/pf0000379949