

Estrategias de difusión de la actividad investigadora: el poder de la **ciencia abierta**

Wenceslao Arroyo Machado

Departamento de Información y Comunicación

Contenidos

1. Problemas de la ciencia abierta
2. Nuevas posibilidades
3. La tripleta ganadora
4. Buenas prácticas

Objetivos

1. Conocer los problemas y desafíos de los datos abiertos
2. Aprender nuevas prácticas para compartir datos y código



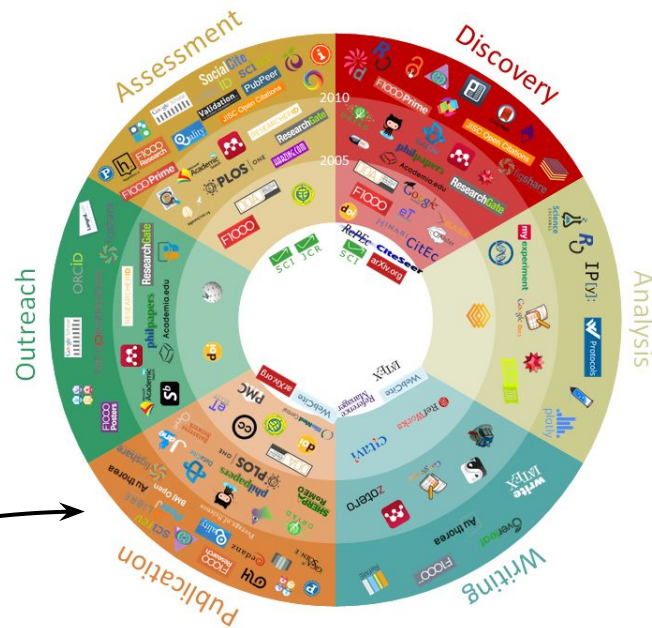
1

Problemas de la ciencia abierta

Consideraciones previas

Problemas de la ciencia abierta

Una parte fundamental del proceso de investigación está en **compartir**, no solo los resultados de investigación sino también los datos y procesos involucrados en ello.



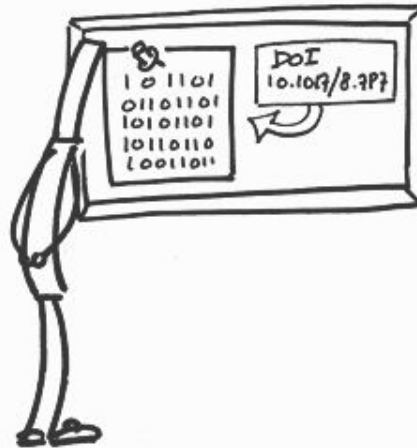
HAY UNA AMPLIA VARIEDAD
DE HERRAMIENTAS

Problemas de la ciencia abierta

FAIR DATA PRINCIPLES



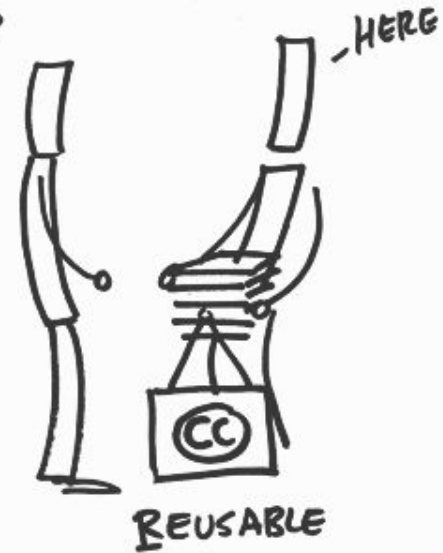
FINDABLE



ACCESIBLE



INTEROPERABLE



REUSABLE

Bezjak S, Clyburne-Sherin A., Conzett P., Fernandes P., Görögh E., Helbig K., ... Heller L.. (2018). Open Science Training Handbook (Version 1.0). Zenodo. <https://doi.org/10.5281/zenodo.1212496>

Problemas de la ciencia abierta

**Datos
Código**

Disponibilidad

Usar repositorios abiertos, con un identificador asociado y licencia que garantice su reutilización

**Datos
Código**

Formato

Emplear formatos abiertos y estandarizados

**Datos
Código**

Documentación

Detallar y explicar qué se incluye exactamente y los procesos

**Datos
Código**

Limpieza

Simplificar y eliminar cualquier tipo de ruido

Código

Reproducibilidad

Asegurar su correcta reproducción

Problemas de la ciencia abierta

De todos ellos los principales desafíos son

Disponibilidad

Uf. Tenemos problemas para encontrar ese sitio.

No podemos conectar al servidor en www.datosdeinvestigacion.com.

Si esa dirección es correcta, aquí hay otras tres cosas que puede probar:

- Vuelva a intentarlo más tarde.
- Compruebe su conexión de red.
- Si está conectado a través de un cortafuegos, compruebe que Firefox tiene permiso para acceder a la web.

[Reintentar](#)

Reproducibilidad

```
> analisisDatos(datos = 'carpeta')
ERROR FATAL
Es necesario un paquete que ya no existe
La versión instalada del software no es la requerida
Warning message:
In analisisDatos(datos = "carpeta") : Falta un archivo
> |
```

Problemas de la ciencia abierta | Disponibilidad

Aunque es una práctica cada vez más extendida y requerida, de manera general todavía no cuenta con la atención y reconocimiento que debe.

COMMENT | 04 June 2019 | Correction [05 June 2019](#)

Make scientific data FAIR

All disciplines should follow the geosciences and demand best practice for publishing and sharing data, argue Shelley Stall and colleagues.

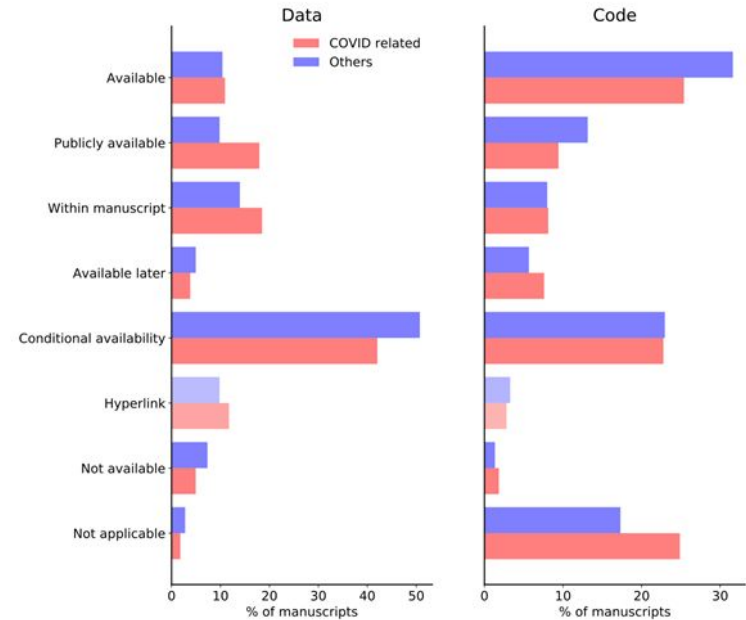
[Shelley Stall](#) , [Lynn Yarmey](#), [Joel Cutcher-Gershenfeld](#), [Brooks Hanson](#), [Kerstin Lehnert](#), [Brian Nosek](#), [Mark Parsons](#), [Erin Robinson](#) & [Lesley Wyborn](#)

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., Parsons, M., Robinson, E., & Wyborn, L. (2019). Make scientific data FAIR. *Nature*, 570(7759), 27–29. <https://doi.org/10.1038/d41586-019-01720-7>

Problemas de la ciencia abierta | Disponibilidad

La pandemia de COVID-19 ha evidenciado la falta de transparencia y compromiso en este sentido.

Los esfuerzos para compartir datos y códigos no han cambiado y siguen siendo minoría.



Larregue, J., Vincent-Lamarre, P., Lebaron, F., & Larivière, V. (2020). COVID-19: Where is the data? *Impact of Social Sciences*.
<https://blogs.lse.ac.uk/impactofsocialsciences/2020/11/30/covid-19-where-is-the-data/>

Problemas de la ciencia abierta | Disponibilidad

Compartir los datos implica no solo hacerlos **públicamente** disponibles sino seguir unas buenas prácticas que garanticen su disponibilidad y correcta reutilización.

A diferencia de los manuscritos, las normas para publicar los datos son generales y fácilmente aplicables.

Data availability

Please provide a Data Availability statement in the Methods section under "Data Availability"; detailed guidance can be found in our [data availability and data citations policy](#). Certain data types must be deposited in an appropriate public structured data depository (details are available [here](#)), and the accession number(s) provided in the manuscript. Full access is required at publication. Should full access to data be required for peer review, authors must provide it.

We encourage provision of other source data in unstructured public depositories such as [Dryad](#) or [figshare](#), or as supplementary information. To maximize data reuse, we encourage publication of detailed descriptions of datasets in [Scientific Data](#).

Fuente: <https://www.nature.com/nature/for-authors/initial-submission>

Problemas de la ciencia abierta | Reproducibilidad

No es suficiente con compartir los datos en bruto y el *script* en el que se ha realizado el análisis.

Es necesario garantizar que estos sean fácilmente **comprensibles** y que el proceso al completo pueda **reproducirse** sin problema.



Fuente: <https://hackernoon.com/its-time-we-code-in-english-e02df6b62ecc>

Problemas de la ciencia abierta | Reproducibilidad

SOLO UNA PEQUEÑA PARTE DE LOS
RESULTADOS SON REPRODUCIBLES



PNAS

An empirical analysis of journal policy effectiveness for computational reproducibility

Victoria Stodden^{a,1}, Jennifer Seiler^b, and Zhaojun Ma^b

^aSchool of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL 61820; and ^bDepartment of Statistics, Columbia University, New York, NY 10027

Edited by David B. Allison, Indiana University Bloomington, Bloomington, IN, and accepted by Editorial Board Member Susan T. Fiske January 9, 2018 (received for review July 11, 2017)

A key component of scientific communication is sufficient information for other researchers in the field to reproduce published findings. For computational and data-enabled research, this has often been interpreted to mean making available the raw data from which results were generated, the computer code that generated the findings, and any additional information needed such as workflows and input parameters. Many journals are revising author guidelines to include data and code availability. This work evaluates the effectiveness of journal policy that requires the data and code necessary for reproducibility be made available postpublication by the authors upon request. We assess the effectiveness of such a policy by (i) requesting data and code from authors and (ii) attempting replication of the published findings. We chose a random sample of 204 scientific papers published in the journal *Science* after the implementation of their policy in February 2011. We found that we were able to obtain artifacts from 44% of our sample and were able to reproduce the findings for 26%. We find this policy—author remission of data and code postpublication upon request—an improvement over no policy, but currently insufficient for reproducibility.

reproducible research | data access | code access | reproducibility policy | open science

putational reproducibility of published results. We use a survey instrument to test the availability of data and code for articles published in *Science* in 2011–2012. We then use the scientific communication standards from the 2012 Institute for Computational and Experimental Research in Mathematics (ICERM) workshop report to evaluate the reproducibility of articles for which artifacts were made available (11). We then assess the impact of the policy change directly, by examining articles published in *Science* in 2009–2010 and comparing artifact ability to our postpolicy sample from 2011–2012. Finally, we discuss possible improvements to journal policies for enabling reproducible computational research in light of our results.

Results

We emailed corresponding authors in our sample to request the data and code associated with their articles and attempted to replicate the findings from a randomly chosen subset of the articles for which we received artifacts. We estimate the artifact recovery rate to be 44% with a 95% bootstrap confidence interval of the proportion [0.36, 0.50], and we estimate the replication rate to be 26% with a 95% bootstrap confidence interval [0.20, 0.32].

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11), 2584–2589. <https://doi.org/10.1073/pnas.1708290115>





2

Nuevas posibilidades

Preparación de los materiales

Nuevas posibilidades

Existen una amplia variedad de posibilidades para llevar a cabo la difusión de estos resultados.

El punto en común en buena parte de ellos está en el uso de notebooks.

communications physics

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [communications physics](#) > [comment](#) > [article](#)

[Comment](#) | [Open Access](#) | [Published: 19 August 2020](#)

Creating an executable paper is a journey through Open Science

[Jana Lasser](#) 

[Communications Physics](#) **3**, Article number: 143 (2020) | [Cite this article](#)

6234 Accesses | **3** Citations | **74** Altmetric | [Metrics](#)

Executable papers take transparency and openness in research communication one step further. In this comment, an early career researcher reports her experience of creating an executable paper as a journey through Open Science.

Lasser, J. (2020). Creating an executable paper is a journey through Open Science. *Communications Physics*, 3(1), 143. <https://doi.org/10.1038/s42005-020-00403-4>

Nuevas posibilidades | Notebooks

Aunque las notebooks nacieron en los 80s, su uso se ha popularizado con el auge de la ciencia de datos.

Integran en una misma interfaz **procesador de textos** (habitualmente en Markdown) y **programación**.

ALGUNAS PERMITEN LA INTERACCIÓN

The Lorenz Differential Equations

Before we start, we import some preliminary libraries. We will also import (below) the accompanying `lorenz.py` file, which contains the actual solver and plotting routine.

```
[1]: %matplotlib inline
from ipywidgets import interactive, fixed
```

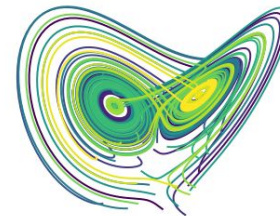
We explore the Lorenz system of differential equations:

$$\begin{aligned}\dot{x} &= \sigma(y - x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= -\beta z + xy\end{aligned}$$

Let's change (σ, β, ρ) with ipywidgets and examine the trajectories.

```
[2]: from lorenz import solve_lorenz
w=interactive(solve_lorenz,sigma=(0.0,50.0),rho=(0.0,50.0))
w
```

sigma 10.00
beta 2.67
rho 28.00



TAMBIÉN SE MUESTRAN LOS GRÁFICOS

Nuevas posibilidades | Notebooks

Destaca sobre todo **Jupyter**, un proyecto nacido en 2014 para facilitar el desarrollo de software de código abierto y la programación interactiva.

Ofrece soporte para aproximadamente **40 lenguajes**, entre los que destacan Python, R y Julia.

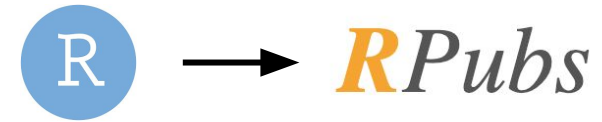


Nuevas posibilidades | Notebooks

Las notebooks pueden compartirse en la red por medio de diferentes plataformas que permiten almacenarlas y/o visualizarlas correctamente.

No ofrece interacción pero hace posible su lectura.

Repositorio de notebooks



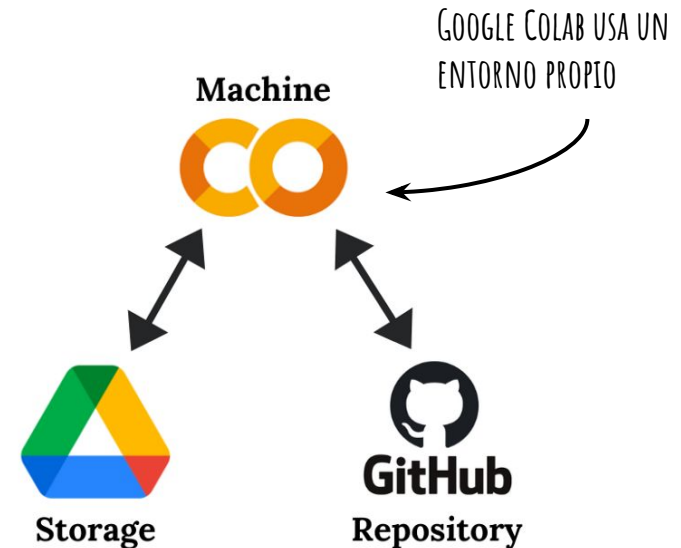
Visualizador de notebooks



Nuevas posibilidades | Entornos colaborativos

Es posible además trabajar directamente con código de manera **colaborativa** haciendo uso de notebooks.

Google Colab es una de las herramientas más populares. Permite integrar Google Drive y GitHub.



Fuente: <https://medium.com/analytics-vidhya/how-to-use-google-colab-with-github-via-google-drive-68efb23a42d>

Nuevas posibilidades | Entornos colaborativos

Aunque se trata de una opción gratuita, ofrece **recursos limitados**.

Existe una versión de pago así como un amplia variedad de alternativas.



**Amazon
SageMaker**



SaturnCloud

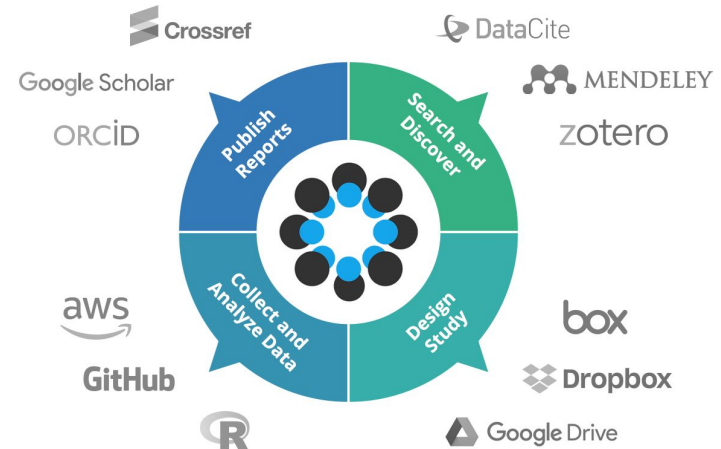
kaggle

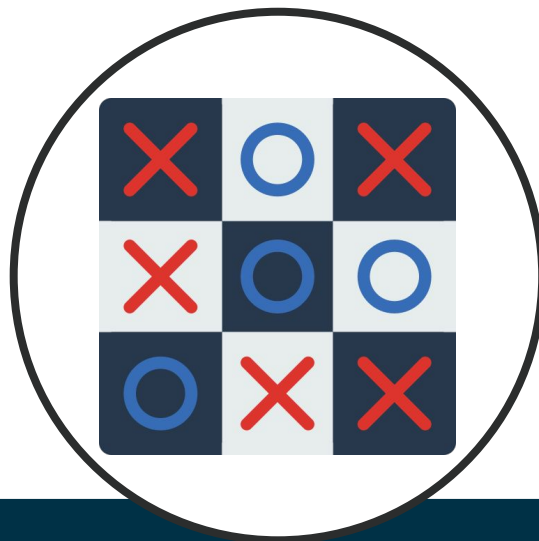


Azure Notebooks

Nuevas posibilidades | Entornos colaborativos

Otra posibilidad se encuentran en herramientas como OSF.io, que ofrece un completo entorno para la investigación, pudiendo integrar numerosas herramientas y repositorios.





3

La tripleta ganadora

Asegurando la disponibilidad y reproducibilidad

La tripleta ganadora | Notebooks

No es suficiente con compartir una notebook. Ciertos aspectos pueden interferir en la correcta ejecución de dichos resultados, como dependencias, versiones concretas de paquetes...

Es necesario ofrecer un entorno concreto para su ejecución.



La tripleta ganadora | Tripleta

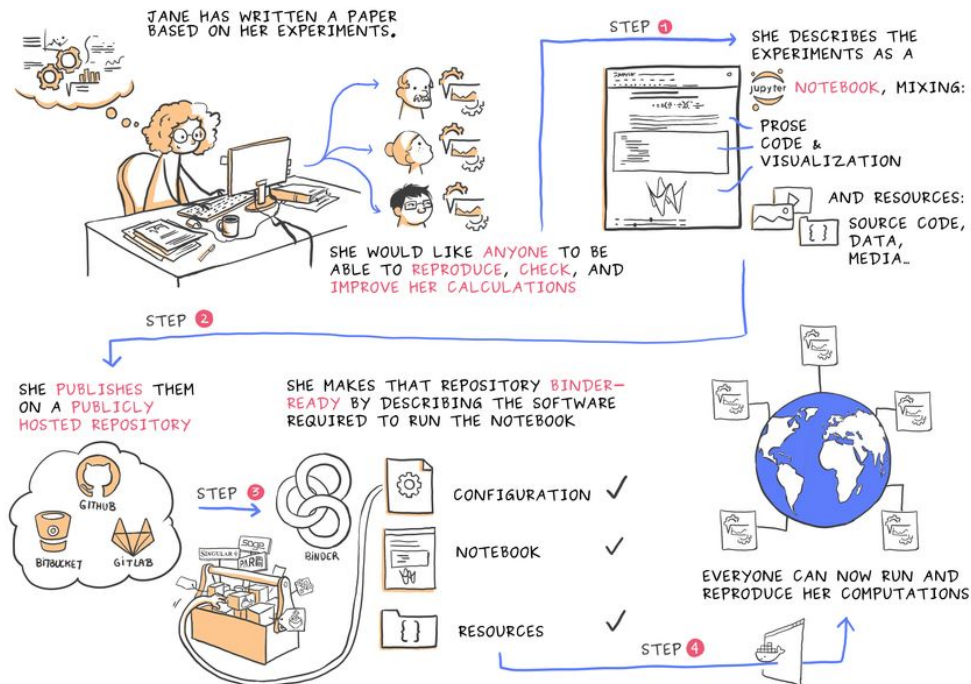
Existen herramientas que permiten partir de repositorios con código y datos y crear todo un entorno con el que ejecutarlo.

Crean **contenedores** que incluyen la configuración exacta para el correcto desarrollo de la aplicación.



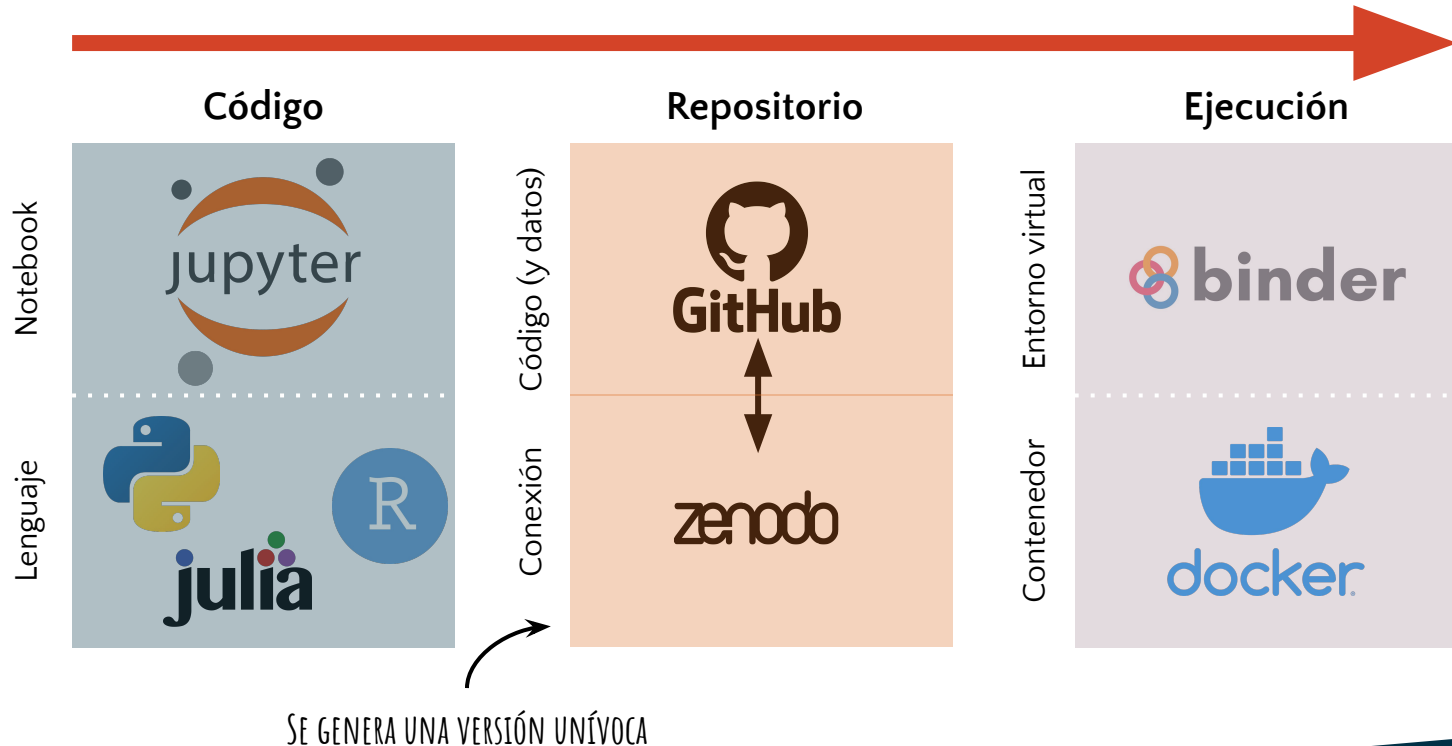
Fuente: <https://bibliometriaobarbarie.com/2019/11/22/trifuerza-open-science/>

La tripleta ganadora | Tripleta



Beg, M., Taka, J., Kluyver, T., Konovalov, A., Ragan-Kelley, M., Thiéry, N. M., & Fangohr, H. (2021). Using Jupyter for reproducible scientific workflows. *Computing in Science & Engineering*, 23(2), 36–46.
<https://doi.org/10.1109/MCSE.2021.3052101>

La tripleta ganadora | Tripleta





4

Buenas prácticas

Asegurando la disponibilidad y reproducibilidad

Buenas prácticas

1. Estructura correctamente el proyecto

Empieza trabajando con los entornos tradicionales y mediante scripts. Una vez estés familiarizado y tengas preparado el análisis pasa a la notebook.

```
datos/  
├── archivos/  
│   ├── analisis.py  
│   ├── archivo.txt  
│   └── archivo(1).txt  
├── datos/  
│   └── sin_nombre.txt  
├── script.py  
├── script(2)_final.py  
└── script_final.py
```

Buenas prácticas

Aprende git

build passing

Repositorio con el material para el libro *Aprende git*. Se escribió originalmente usando *Markdown* en su versión *Kramdown*, aunque finalmente hemos encontrado que funciona mejor transformándolo con *Pandoc*.

Te lo puedes *descargar* en cualquiera de las versiones en *ePub*, *comprar* en Amazon en formato *ebook* o en formato *físico*. Si necesitas generar otro formato, puedes inspirarte en los conversores que hay en el directorio *utils*.

Índice

1. Introducción.
2. Uso básico.
3. Resolviendo conflictos y otros problemas.
4. Flujos de trabajo y otras buenas prácticas.
5. Trabajando en GitHub.
6. Trabajando con *hooks* y otros temas de fontanería.

Fuente: <https://github.com/II/aprende-git>

2. Usa control de cambios

Si tienes previsto usar GitHub para difundir el código, comienza desde el principio a usarlo. El control de cambios mediante git ofrece muchas ventajas.

Buenas prácticas

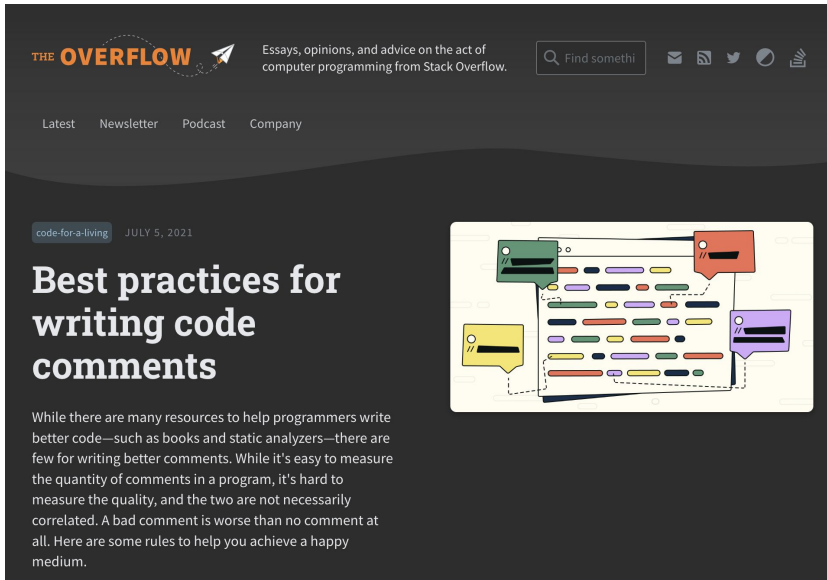
3. No empieces desde una notebook

Empieza trabajando con los entornos tradicionales y mediante scripts. Una vez estés familiarizado y tengas preparado el análisis pasa a la notebook.



Fuente: <https://towardsdatascience.com/how-and-why-to-share-scientific-code-64fbd385a67>

Buenas prácticas



Fuente: <https://stackoverflow.blog/2021/07/05/best-practices-for-writing-code-comments/>

4. Explica todo bien

Desde el comienzo realiza anotaciones, tanto dentro como fuera del código. Puedes usar un pequeño documento a modo de guía para identificar correctamente todos los archivos y pasos.

No olvides los archivos README.

Buenas prácticas

5. No te compliques innecesariamente

Si ya existe un paquete que hace lo que necesitas o si un repositorio o lenguaje se acomodan perfectamente a tu problema, no hagas cambios al respecto.

Repository Name	Information on fees/costs	Size limits	Integrated with <i>Scientific Data's</i> manuscript submission system	Re3data / FAIRsharing entry
Dryad Digital Repository	\$120 USD for first 20 GB, and \$50 USD for each additional 10 GB	None stated	Yes ✓	view FAIRsharing entry
figshare	100 GB free per <i>Scientific Data</i> manuscript. Additional fees apply for larger datasets	1 TB per dataset	Yes ✓ - To qualify for the 100 GB of free storage, data must be uploaded to figshare via our submission system. Download instructions.	view FAIRsharing entry
Harvard Dataverse	Contact repository for datasets over 1 TB	2.5 GB per file, 10 GB per dataset	No	view re3data entry
Open Science Framework	Free of charge	5 GB per file, multiple files can be uploaded	No	view FAIRsharing entry
Zenodo	Donations towards sustainability encouraged	50 GB per dataset	No	view re3data entry
Science Data Bank	Free of charge	8 GB per file, no limit to dataset size	No	view FAIRsharing entry

Fuente: <https://www.nature.com/sdata/policies/repositories>

Buenas prácticas

Publication date:

October 28, 2020

DOI:

DOI [10.5281/zenodo.4148941](https://doi.org/10.5281/zenodo.4148941)

Keyword(s):

altmetrics

Twitter

Library and Information Science

Microbiology

Social Network Analysis

Related identifiers:

Compiled by

https://github.com/Wences91/social_media_communities (Other)

Source of

[10.1007/s11192-021-04167-8](https://doi.org/10.1007/s11192-021-04167-8) (Journal article)
[10481/71230](https://doi.org/10.1007/s11192-021-04167-8) (Preprint)

Supplementary material

[10.5281/zenodo.4332921](https://doi.org/10.5281/zenodo.4332921) (Other)

License (for files):

 Creative Commons Attribution 4.0 International

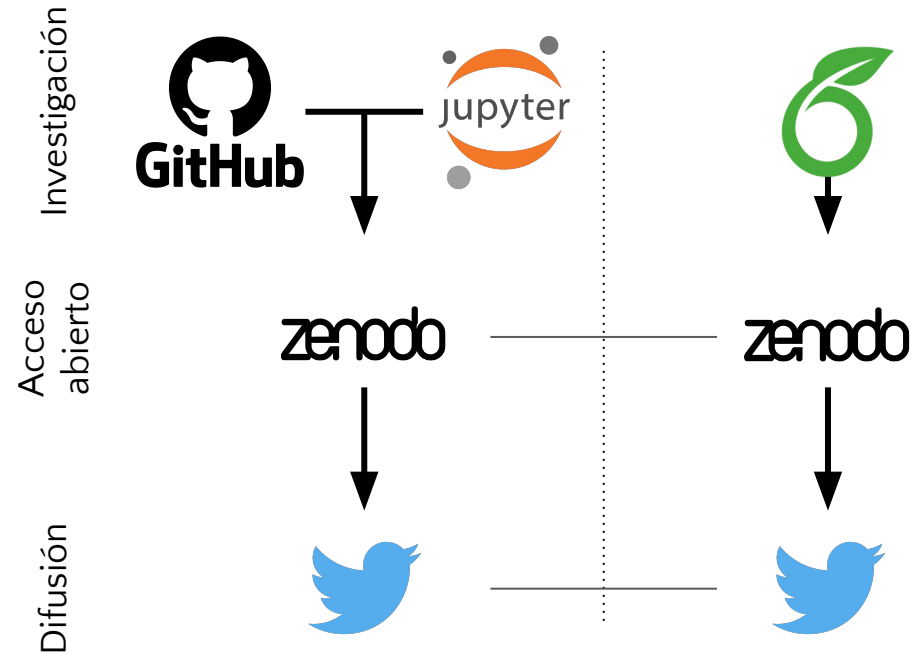
6. Enlaza todos los recursos

Conecta todos los resultados de investigación que estén bajo un mismo proyecto. Los repositorios permiten establecer dichas conexiones.

Buenas prácticas

7. Lleva a cabo un plan de difusión

Empieza trabajando con los entornos tradicionales y mediante scripts. Una vez estés familiarizado y tengas preparado el análisis pasa a la notebook.





¡Muchas gracias!

¿Alguna pregunta?

zenodo

Presentación disponible en <https://doi.org/10.5281/zenodo.5713266>

