Datos abiertos: Desde el acceso abierto al producto de investigación, a la ciencia abierta como proceso

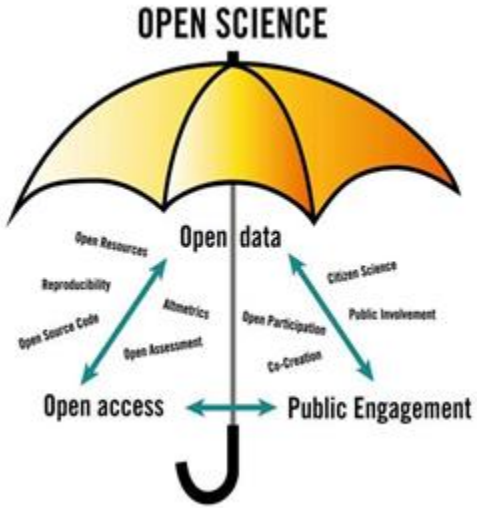
Lucía Castillo Iglesias
Universidad de Concepción
MiNSoL

Contexto: Ciencia abierta

El concepto de **ciencia abierta** sirve como paraguas para referirnos al conjunto de prácticas, políticas y principios de apertura, transparencia y democratización del conocimiento.

Contexto de crisis de la credibilidad del conocimiento científico: ***reproducibility / replicability crisis***

Pandemia COVID como ejemplo del potencial de las prácticas científicas abiertas



Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding

Roujian Lu*, Xiang Zhao*, Juan Li*, Peihua Niu*, Bo Yang*, Honglong Wu*, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, Yuhai Bi, Xuejun Ma, Faxian Zhan, Liang Wang, Tao Hu, Hong Zhou, Zhenhong Hu, Weimin Zhou, Li Zhao, Jing Chen, Yao Meng, Ji Wang, Yang Lin, Jianying Yuan, Zhihao Xie, Jinmin Ma, William J Liu, Dayan Wang, Wenbo Xu, Edward C Holmes, George F Gao, Guizhen Wu¶, Weijun Chen¶, Weifeng Shi¶, Wenjie Tan¶

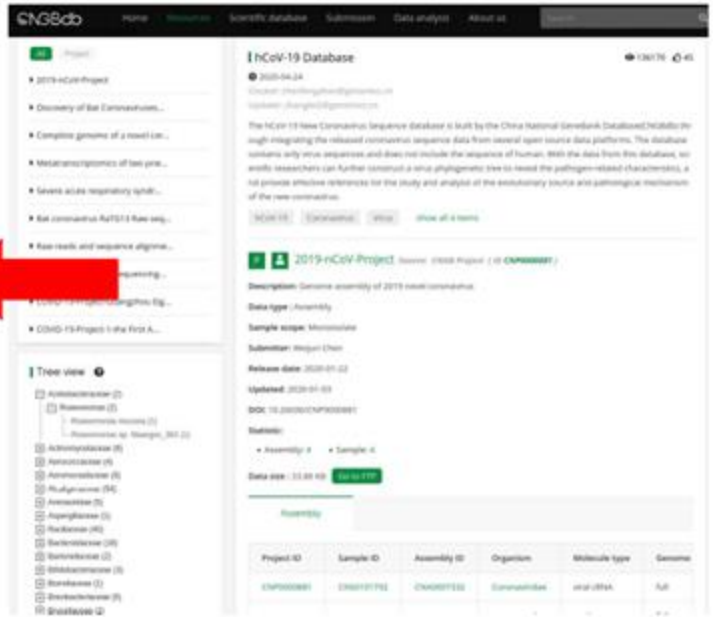
Summary

Background In late December, 2019, patients presenting with viral pneumonia due to an unidentified microbial agent were reported in Wuhan, China. A novel coronavirus was subsequently identified as the causative pathogen, provisionally named 2019 novel coronavirus (2019-nCoV). As of Jan 26, 2020, more than 2000 cases of 2019-nCoV infection have been confirmed, most of which involved people living in or visiting Wuhan, and human-to-human transmission has been confirmed.

Lancet 2020; 395: 565–74
Published Online
January 29, 2020
https://doi.org/10.1016/S0140-6736(20)30251-8
*Contributed equally

Results

From the nine patients' samples analysed, eight complete and two partial genome sequences of 2019-nCoV were obtained. These data have been deposited in the China National Microbiological Data Center (accession number NMDC10013002 and genome accession numbers NMDC60013002-01 to NMDC60013002-10) and the data from BGI have been deposited in the China National GeneBank (accession numbers CNA0007332–35).



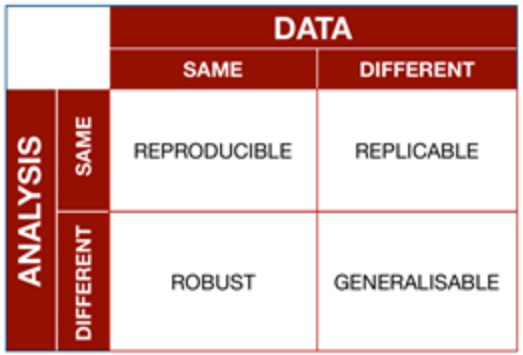
¿Por qué datos abiertos?

Crisis de reproducibilidad

Reproducible: Misma pregunta, mismos análisis, mismos datos (¿debería ser trivial o no?)

Necesidad de EXPLICITAR todas las decisiones tomadas en el proceso

Machine-readable data: ¿quieres ser considerado en los meta-análisis de tu área?



Políticas editoriales (ej. Nature, Science, PLoS, **JML**)

"We require articles published in the Journal to make publicly available any stimuli, data, analysis code, and computational models associated with the research. Because these materials are an important part of the research that the Journal is reviewing, we require authors to provide a (private) link to them at the point of submission. Please include this link on the title page of your manuscript. Manuscripts that do not include access to these materials will usually be returned to authors."

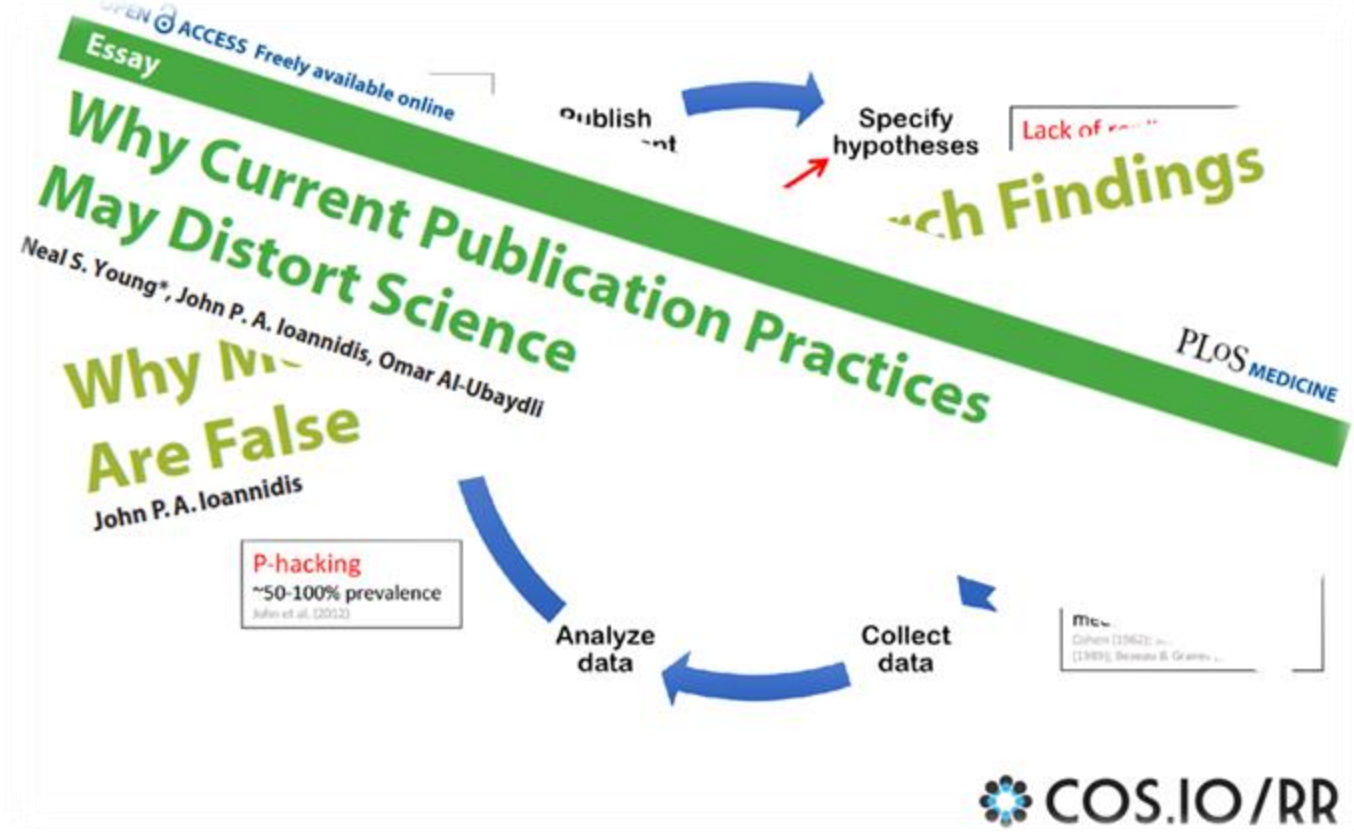
Crisis de reproducibilidad/replicabilidad

COS.IO/RR

Crisis de reproducibilidad/replicabilidad

COS.IO/RR

Crisis de reproducibilidad/replicabilidad



Open Science Collaboration (2015), "Estimating the reproducibility of psychological science"

- Consorcio de laboratorios intentó reproducir 100 artículos publicados en revistas de alto impacto

- 97% de los artículos originales con p < .05
- 36% de las réplicas con p< .05

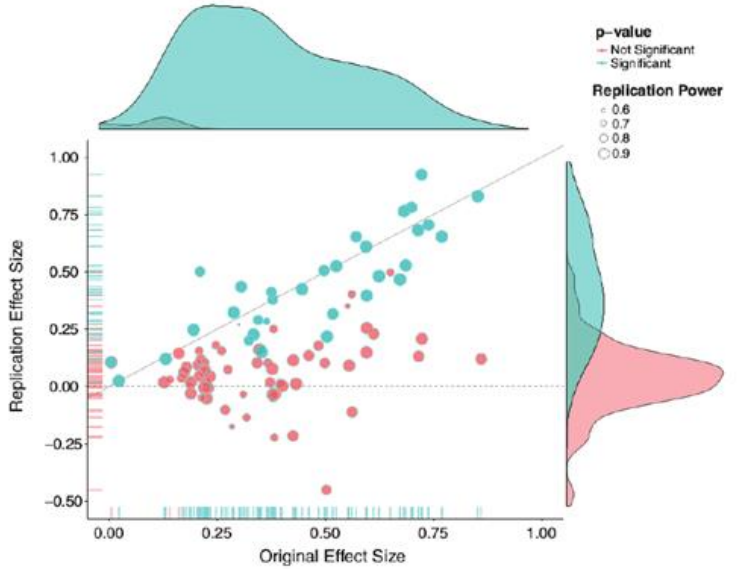
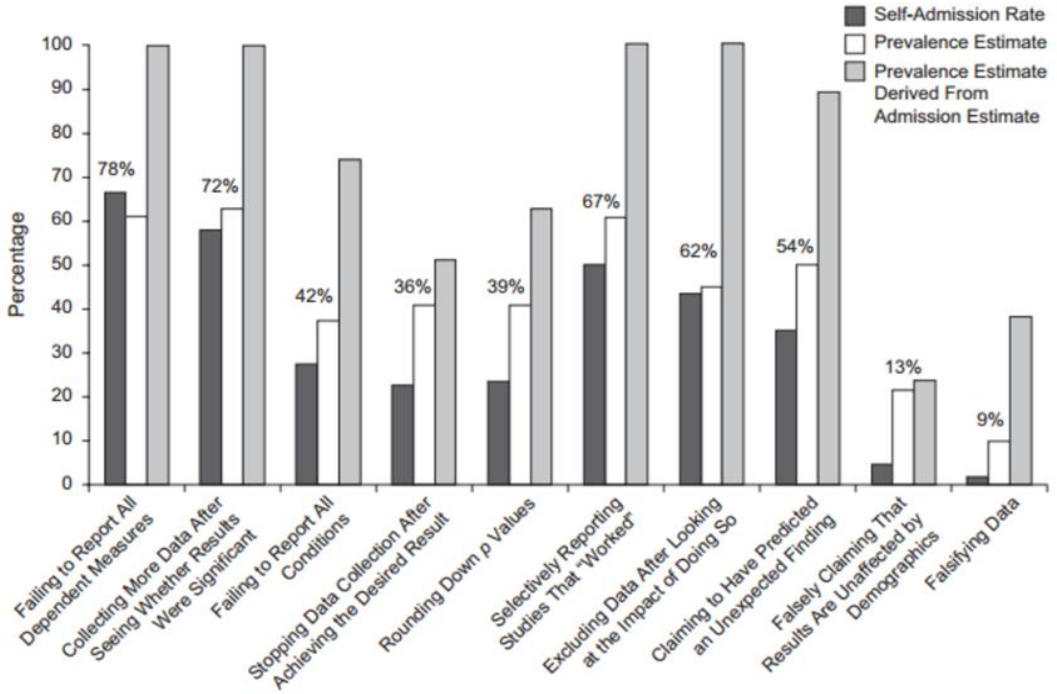


Fig. 3. Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Manipulación de datos

John, Loewenstein, & Prelec (2012)



Contexto: Política de ciencia abierta ANID

Vigente desde 2022
Alineada con objetivos UNESCO
Construcción de una Infraestructura Nacional de Acceso a la Información Científica

SciELO Chile
Acceso abierto a textos completos de revistas científicas chilenas.

BEIC
Acceso gratuito a revistas en texto completo.

Repositorio ANID
Resultados de investigación financiados por ANID.

Portal del Investigador
Perfiles de investigadores y su productividad.

DataCiencia
Producción científica chilena.

+ 5 proyectos 2023

Comisión Europea publica informe en el que se destaca el trabajo de Chile en torno a normativas vigentes sobre acceso abierto al conocimiento

El documento elaborado por Pilar Rico-Castro y Laura Bonora, profesionales de la Unidad de Acceso Abierto, Repositorios y Revistas de la FECYT (España), expone el panorama de América Latina, el Caribe y la Unión Europea en políticas de acceso abierto y genera un espacio para avanzar en el diálogo político en torno a estas prácticas y legislaciones.

Descargar informe

La **Política de Acceso Abierto a la Información Científica y a Datos de Investigación** financiados con fondos públicos de la ANID es un esfuerzo de la **Agencia Nacional de Investigación y Desarrollo** para dejar disponibles en el Repositorio Institucional todos los resultados de las investigaciones financiadas con recursos estatales. Entró en vigencia en 2022 para algunos instrumentos de la Agencia y luego de una evaluación, se extenderá

Contexto: Política de datos ANID

Política de Acceso Abierto ANID vigente desde mayo 2022 *para todos aquellos instrumentos que produzcan resultados científicos*

- **Presentar un plan de gestión de datos**, al inicio y cierre del proyecto, y cada vez que los beneficiarios entreguen a la ANID los informes parciales.
- **Depositar los datos de investigación en el repositorio** de la ANID, en repositorios disciplinarios o institucionales, en el plazo de un año después de presentado el informe final.
- Si los datos de investigación se depositaron en repositorios distintos al de la ANID, el beneficiario debe informar a la contraparte ANID, la URL o los identificadores persistentes de dichos datos.
- Si los datos de investigación están afectos a confidencialidad /acuerdos previos con terceros, el beneficiario debe proporcionar acceso público a los metadatos y entregar acceso abierto al contenido cuando cesen esas obligaciones.

Contexto: Plan de gestión de datos ANID

ANID solicita completar ocho campos, tres de ellos con información administrativa. En algunos habrá lista desplegable de opciones:

1. Proyecto: número de folio, nombre director(a).
2. Identificador ORCID del director(a).
3. Fecha de presentación /actualización del plan.
4. Tipos de datos que serán resultados de la investigación.
5. Tamaño estimado del set de datos.
6. Procedencia de los datos.
7. URL del repositorio donde se depositarán los datos.
8. Identificador único persistente para encontrar los datos almacenados en un repositorio.

¿Qué entendemos por datos?

En un sentido global, ***todo resultado del proceso de investigación:*** archivos de pre o post procesamiento de software, base de datos, cuadernos de campo o anotaciones de laboratorio, cuestionarios o transcripciones, documentos de texto electrónicos, fotografías o películas, algoritmos, planillas de datos, video, entre otros.

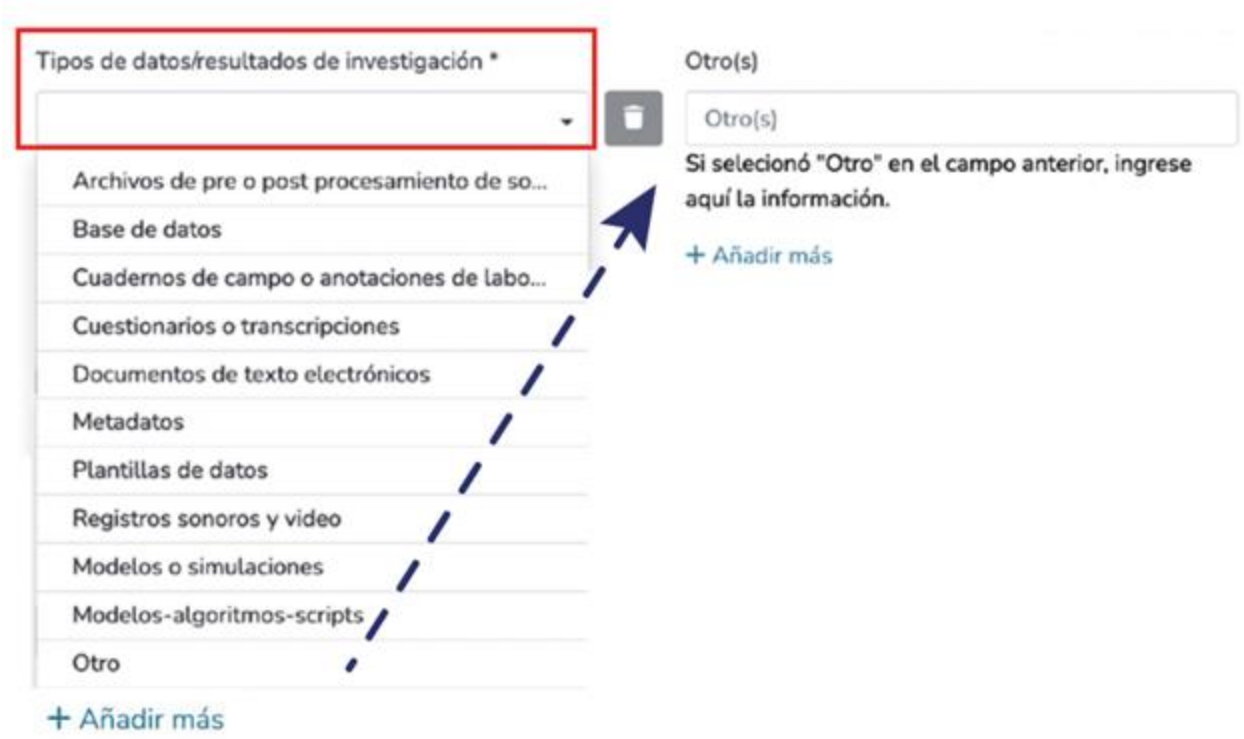
En ciencias humanas experimentales, más frecuentemente:

- **Datos experimentales (.csv, .eeg, etc.)**
- **Datos de participantes (cuestionarios, información demográfica, etc.)**
- **Protocolos y/o registros experimentales**

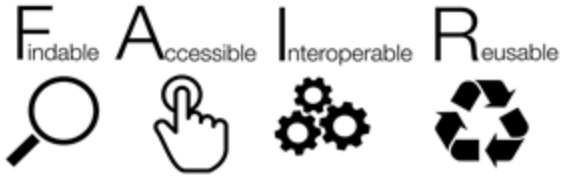
Cuestiones clave:

- Separación entre datos identificables y experimentales
- Concepto de **metadato**

¿Qué entiende ANID por datos?



¿Qué entendemos por *meta*dat



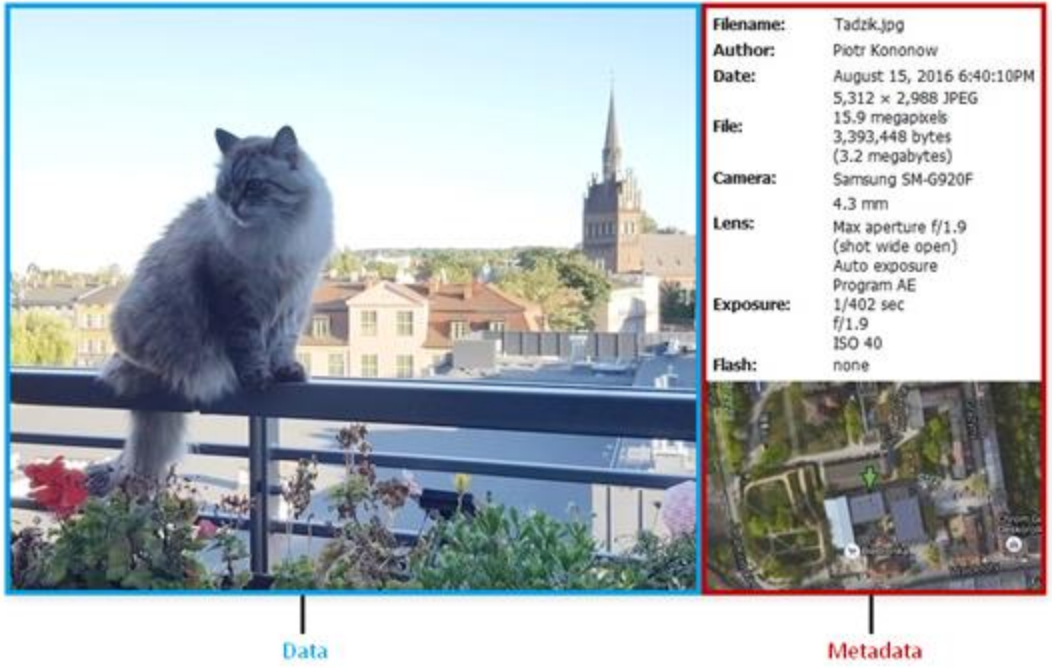
Para que los datos sean “encontrables” e “interoperables”, se requiere información respecto de lo que contienen

El principal estándar de metadatos es Dublin Core, que define un conjunto de atributos mínimos (título, resumen, autores, materias, fechas, etc.) que debieran describir cada documento de un repositorio.

Estándar ANID: Protocolo OAI-PMH (Dublin Core XML)

Protocolo de metadatos abiertos que permite la recopilación automática de metadatos de repositorios digitales. Se utiliza para compartir y acceder a información y contenido digital en un formato estándar, lo que facilita la integración y el intercambio de información entre diferentes repositorios y sistemas.

¿Qué entendemos por *meta*datos?



Pero…

Plan de manejo de datos

Idealmente, generado antes de comenzar la fase experimental del proyecto

Documento VIVO / DINÁMICO

Considera, entre otros:

- Qué datos se van a registrar y cómo
- Cómo van a documentarse esos datos
- Aspectos éticos y legales (identificación y otros) /*desidentificación*
- Almacenaje y respaldo
- Preservación de datos (post término del proyecto)
- Disponibilidad de datos
- Responsabilidades

Plan de manejo de datos: Ejemplo

Proyecto: experimento EEG

1. **Qué datos se van a registrar y cómo**

   Tres tipos de datos: **datos de EEG, datos demográficos de los participantes, datos de contacto de los participantes**

   Cómo:

- Datos de EEG se registran directamente a partir del registro de actividad cerebral mediante EEG. El sistema los guarda en formato .eeg. Adicionalmente se registran datos de texto en formato .vmrk y .vhdr. Los datos de ,eeg se pre-procesan con el software BrainVision Analizer o equivalente y se guardan en formato .csv para su análisis en R.
- Datos demográficos se registran a través de un formulario de Google, que incluye edad, sexo, nivel socioeconómico, nivel educacional. El formulario ordena estos datos y les asigna un código de 4 dígitos para identificación del participante, en una planilla .csv
- Datos de contacto se registran manualmente en un formulario de ingreso, el que no es digitalizado. Estos datos se guardan de acuerdo a las condiciones definidas por el Comité de Ética, asegurando que no pueda establecerse ninguna conexión directa entre los datos experimentales y de contacto.

Plan de manejo de datos: Documentación

¿Qué necesita ser documentado? Tres niveles:

- **Proyecto**
- **Archivo**
- **Variable**

The citation advantage of linking publications to research data

Giovanni Colavizza, Iain Hrynaszkiewicz, Isla Staden, Kirstie Whitaker, Barbara McGillivray

Published: April 22, 2020 • https://doi.org/10.1371/journal.pone.0230416

¿Para qué?

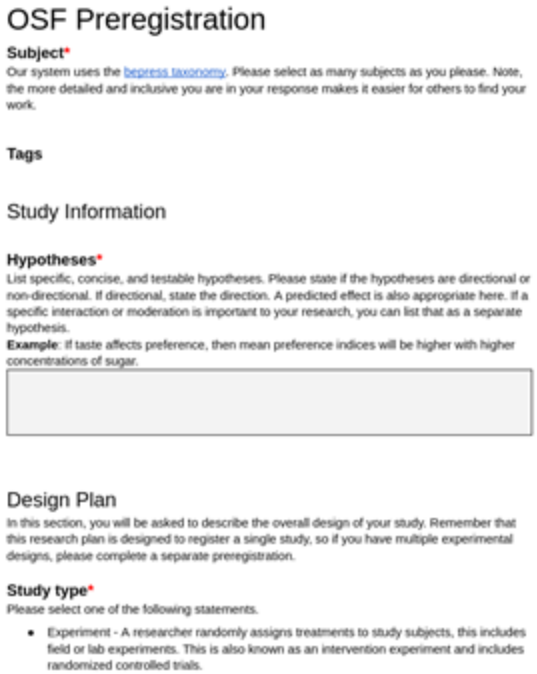
Para lograr la **autonomía de los datos**: para los demás (otros investigadores), para tu inserción en la disciplina (estandarización), pero también para tu ***"yo del futuro"*** y para otros miembros de tu laboratorio (estudiantes, colaboradores)

Plan de manejo de datos: Documentación

¿Qué requiere ser documentado?

Básicamente, todo lo que puedas

- **Proyecto:** todo lo que incluirías normalmente en un pre-registro: hipótesis, variables, medidas, magnitudes
- Archivo
- Variable



OSF Preregistration

Subject*

Our system uses the bepress taxonomy. Please select as many subjects as you please. Note, the more detailed and inclusive you are in your response makes it easier for others to find your work.

Tags

Study Information

Hypotheses*

List specific, concise, and testable hypotheses. Please state if the hypotheses are directional or non-directional. If directional, state the direction. A predicted effect is also appropriate here. If a specific interaction or moderation is important to your research, you can list that as a separate hypothesis.

Example: If taste affects preference, then mean preference indices will be higher with higher concentrations of sugar.

Design Plan

In this section, you will be asked to describe the overall design of your study. Remember that this research plan is designed to register a single study, so if you have multiple experimental designs, please complete a separate preregistration.

Study type*

Please select one of the following statements.

- Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.

Plan de manejo de datos: Documentación

¿Qué requiere ser documentado?

- Proyecto
- **Archivo:** todo lo requerido para la lectura independiente del archivo
  - **Principios básicos: Estandarización & consistencia**
  - Puede ser entendido como un “mapa” del archivo: ¿qué elementos contiene? ¿cómo se miden? ¿cómo se relacionan entre sí?
  - Incluir los procesos y pre-procesos que dieron origen al archivo (scripts anotados, métodos de anonimización, codificación, etc.)
  - Lo que normalmente se asocia a metadatos: archivos README, archivos de acompañamiento
- Variable

Plan de manejo de datos: Documentación

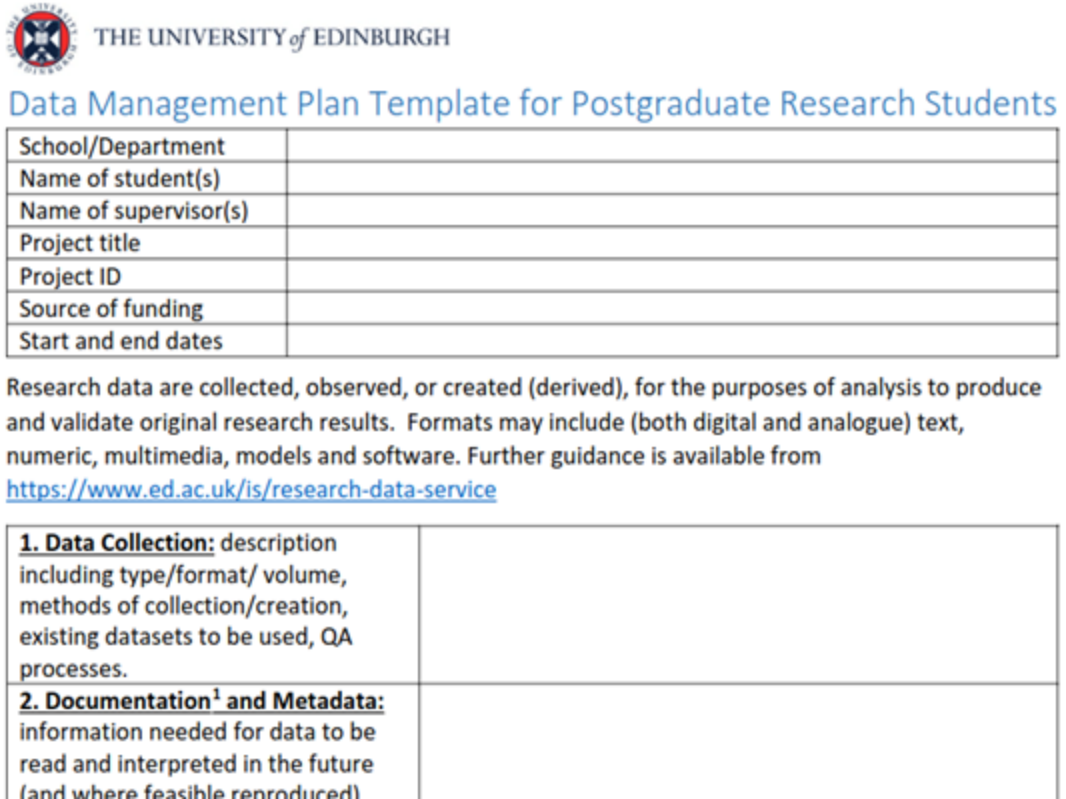
¿Qué requiere ser documentado?

- Proyecto
- Archivo
- **Variable:**
  - Puede ser entendido como un “glosario”: ¿Qué significa “condición” en tu archivo? ¿Qué se entiende por “factor”, qué tipo de dato es (número, vector, lista, etc.), qué niveles tiene y qué significan?
  - Códigos, etiquetas, categorías
  - Codificación de valores faltantes (NA)
  - Explicación de agrupamientos de variables y otras operaciones de post-proceso

Documentación de un proyecto

Data Management Templates

Planes de manejo de datos basados en el proyecto Horizon 2020 de la Unión Europea



THE UNIVERSITY of EDINBURGH

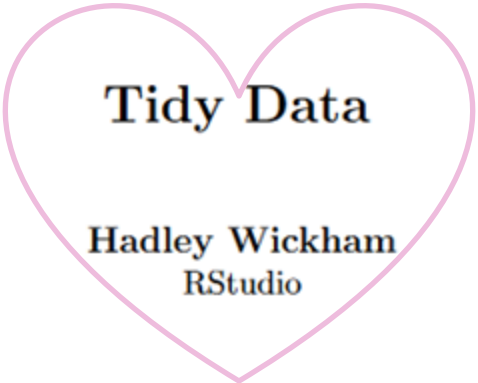
Data Management Plan Template for Postgraduate Research Students

School/Department	
Name of student(s)	
Name of supervisor(s)	
Project title	
Project ID	
Source of funding	
Start and end dates	

Research data are collected, observed, or created (derived), for the purposes of analysis to produce and validate original research results. Formats may include (both digital and analogue) text, numeric, multimedia, models and software. Further guidance is available from https://www.ed.ac.uk/is/research-data-service

1. Data Collection: description including type/format/ volume, methods of collection/creation, existing datasets to be used, QA processes.	
2. Documentation[1] and Metadata: information needed for data to be read and interpreted in the future (and where feasible reproduced)	

Documentación de archivos y variables



Principios básicos: Estandarización y consistencia

¿Archivos, carpetas, variables? Recuerda la metáfora del mapa: necesito saber dónde va esto que estoy mirando en el sistema del proyecto

An Archive.txt file will be created, where information about the project and what is stored in each folder will be created. Importantly, we will create a document where the research design and data collection process are described. The project will have the following folder structure:

1. Background
   a. Literature
   b. Write-up
   c. Notebook
   d. Meetings
2. Experimental-setup
   a. Lab notebook
   b. Materials
   c. Code
3. Raw-Data
   a. Eeg-files
   b. History-files
   c. Export-files
4. Analysis
   a. Processed-data
   b. R-scripts
      i. Data-wrangling
      ii. Data-analysis
      iii. Data-visualization

Folder name	File name(s)	Description
Analysis	behavioural.csv	.csv that contains participants' responses as downloaded from Testable for analysis. This .csv is the output of merge-sess.R and contains all participants' data (i.e., merged by group).
Analysis	data_analysis.R	R script for data analysis done

¿Cuándo documentar?

Desde el día 0

Revisión periódica y ante cualquier cambio

Práctica sistemática, parte del entrenamiento básico del laboratorio: Generación de protocolos y *buenas prácticas* a nivel de lab

Formación de nuevos investigadores que puedan dialogar con el conocimiento de punta de sus disciplinas

¿Cuándo documentar?

Flujo de trabajo:
(0. Determinación de las convenciones y preferencias dentro de la disciplina y laboratorio)

1. **Pre-registro o plan de trabajo**
2. **Plan de manejo de datos (templates)**
3. **Creación de proyecto en OSF o similar**
4. **Uso de notebooks o diario de laboratorio**
5. **Metadatos: Preparación de archivos y variables para su publicación**
6. **Reproducibilidad: Ejecución de análisis a partir únicamente de los datos y códigos publicados**

Journal of Memory and Language 104 (2019) 126–127

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Editorial

New initiatives to promote open science at the *Journal of Memory and La*

Journal of Memory and Language

Volume 125, August 2022, 104332

Toward standard practices for sharing computer code and programs in neuroscience

Stephen J Eglen, Ben Marwick, Yaroslav O Halchenko, Michael Hanke, Shoaib Sufi, Padraig Gleeson, R Angus Silver, Andrew P Davison, Linda Lanyon, Mathew Abrams, Thomas Wachtler, David J Willshaw, Christophe Pouzat & Jean-Baptiste Poline

Nature Neuroscience 20, 770–773 (2017) | Cite this article

16k Accesses | 57 Citations | 304 Altmetric | Metrics

Computational techniques are central in many areas of neuroscience and are relatively easy to share. This paper describes why computer programs underlying scientific publications should be shared and lists simple steps for sharing. Together with ongoing efforts in data sharing, this should aid reproducibility of research.

Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy

Anna Laurinavichyute, Himanshu Yadav, Shravan Vasishth

"We propose two simple steps that can increase the reproducibility of published papers: share the analysis code, and attempt to reproduce one's own analysis using only the shared materials."

¿Cómo y dónde almacenar de manera abierta?

Repositorio ANID

Repositorios institucionales

Repositorios de materiales y datos (Zenodo, Figshare)

Repositorios generales de acceso abierto (OSF)

Interaction promotes adaptation - maze game

588.8KB Make Private Public ₽ 0 ···

Contributors: **Lucia Castillo**, Gregory Mills

Date created: 2019-03-19 09:53 AM | Last Updated: 2019-11-08 09:16 AM
Create DOI
Category: Project

Description:

Coordination between speakers in dialogue requires balancing repetition and change, the old and the new. Interlocutors tend to re-use established forms, relying on communicative precedents. Yet linguistic interaction also necessitates adaptation to changing contexts or dynamic tasks, which might lead to abandoning existing precedents in favor of better communicative alternatives. We explored this tension using a maze game task in which individual participants and interacting pairs had to describe figures and their positions in one of two possible maze types: a regular maze, in which the grid-like structure of the maze is highlighted, and an irregular maze, in which specific parts of the maze are salient. Participants repeated this task several times. Both individuals and interacting pairs were affected by the different maze layouts, initially using more idiosyncratic description schemes for irregular mazes and more systematic schemes for regular mazes. Interacting pairs, but not individuals, abandoned their unsystematic initial descriptions in favor of a more systematic approach, which was better adapted for repeated interaction. Our results show communicative conventions are initially shaped by context, but interaction opens up the possibility for change if better alternatives are available.

Keywords: Convention – Adaptation – Interaction – Alignment – Reference

License: *Add a license*

Wiki

Add important information, links, or images here to describe your project.

Files

Click on a storage provider or drag and drop to upload

Filter

Name	Modified
Interaction promotes adaptation - maze game	
− OSF Storage (United States)	
+ data	
+ scripts	

Citation

Components

Add Component Link Projects

Add components to organize your project.

Tags

Add a tag to enhance discoverability

Recent Activity

Consideraciones finales

Cuestiones fundamentales: sea consistente, sea breve, sea explícito, no confíe en Windows :)

Genere protocolos revisables y compartidos entre los miembros de su laboratorio: la reproducibilidad parte por casa

Control de versiones: clave si trabaja con software! El mismo análisis con los mismos datos puede cambiar si cambia el paquete de R

- En archivos de texto y bases de datos
- En scripts de análisis: GitHub, GitRepository

Licencias: Una licencia abierta estándar asegura la interoperabilidad legal

Consideraciones finales

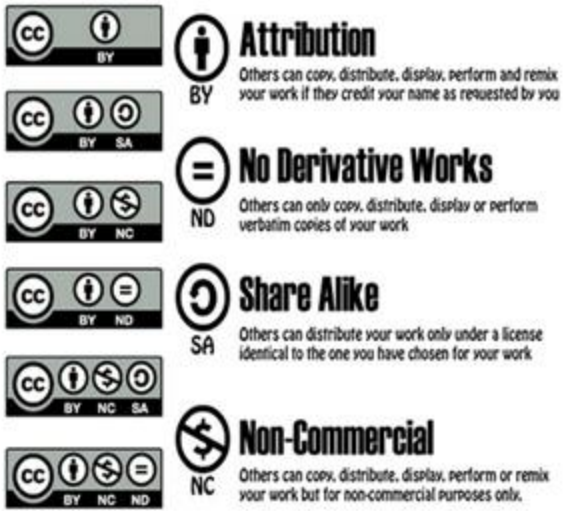
Tipos de licencias (no obligatorio, pero deseable):

CC licenses conformant with the “Open Definition” are:

Public Domain Dedication (CC0 1.0): waives copyright and related rights (e.g. databases).

Attribution (CC-BY-4.0): gives others maximum freedom to reuse (i.e. copy, redistribute, adapt) your work, provided they give appropriate credit.

Attribution Share-Alike (CC-BY-SA-4.0): same as CC-BY-4.0, but requires redistribution of derivative works under this same license.



Links de interés

Manual ANID para el desarrollo del Plan de Gestión de Datos

Generador automático de metadatos Dublin Core

Fairification of data

Nine simple ways to make it easier to (re)use your data (White et al., 2013)

Stanford Psychology Guide to Doing Open Science: Data Sharing


Access for the public
Compliant with grant rules
Practitioners can apply your findings
Taxpayers get value for money
Higher citation rates
Access for reseachers in developing countries
Your research can influence policy
More exposure for your work