

LA-CoNGA Physics 2021
Universidad de los Andes
Mérida - Venezuela

Desarrollar una Infraestructura Sostenible y Reproducible para un grupo pequeño de Investigación

Mildred Arias
Mildarias181@gmail.com

1.- Introducción

Pensar en la reproducibilidad desde el inicio del proyecto es la mejor manera de ahorrar tiempo y aprovechar al máximo las herramientas disponibles. Crear conexiones entre datos, código, metodología, así como diversos colaboradores, puede parecer una tarea abrumadora si no se ha planificado desde el principio. Al documentar y compartir los flujos de trabajo y los procesos del proyecto para la investigación y los investigadores, podemos garantizar la sostenibilidad y la reutilización de la investigación tanto para los desarrolladores como para los futuros usuarios y así evitar la duplicidad del trabajo que ya se hubiera realizado.

El diseño del proyecto para una investigación reproducible abarca una variedad de aspectos, los usuarios esperados o la audiencia objetivo, los recursos disponibles y las habilidades requeridas en el proyecto. También se requiere que los investigadores exploren los posibles resultados, planifiquen para abordar los desafíos o riesgos esperados y aseguren la diversidad de las partes interesadas.

2.- Objetivos

- Adoptar y ampliar los estándares abiertos existentes
- Compartir datos con el fin de fomentar la innovación abierta por parte del grupo de investigación.
- Usar plataformas abiertas existentes, cuando fuera posible, a fin de ayudar a automatizar el intercambio de datos, a conectar herramienta o sistema con otros y agregar flexibilidad para adaptarse a necesidades futuras.
- Invertir en el software como un bien público.
- Desarrollar el nuevo código de software para que sea de fuente abierta, que cualquiera pueda ver, copiar, modificar y compartir, y distribuir el código en repositorios públicos.
- Posibilitar la innovación al compartir libremente sin restricciones, colaborando de manera amplia y co-creando herramientas.

3.- Investigación Abierta

Una investigación abierta, es la práctica de hacer que “los resultados primarios de los resultados de la investigación financiados con fondos públicos (publicaciones y datos de investigación) sean accesibles al público en un formato digital con restricción mínima o nula “. Para lograr esta apertura en la investigación, cada elemento del proceso de investigación debe:

- **Estar disponible públicamente:** Es difícil de usar y beneficiarse del conocimiento oculto detrás de barreras como contraseñas y muros de pago.
- **Ser reutilizables:** Los resultados de la investigación deben tener la licencia adecuada para que los posibles usuarios conozcan las limitaciones de la reutilización.
- **Sea transparente:** Con metadatos apropiados para proporcionar declaraciones claras sobre cómo se produjo el resultado de la investigación y qué contiene.

El proceso de investigación generalmente tiene la siguiente forma: los datos se recopilan y luego se analizan (generalmente mediante software). Este proceso puede implicar el uso de hardware especializado. Luego se publican los resultados de la investigación. A lo largo del proceso, es una buena práctica que los investigadores documenten su trabajo en cuadernos. La investigación abierta tiene como objetivo abrir cada uno de estos elementos:

- **Datos abiertos:** Documentar y compartir datos de investigación abiertamente para su reutilización. Además de poner los recursos educativos a disposición del público para su reutilización y modificación.
- **Software de código abierto:** Documentar el código de investigación y las rutinas, y hacerlos accesibles y disponibles libremente.
- **Hardware abierto:** Documentar diseños, materiales y otra información relevante relacionada con el hardware, y hacerlos accesibles y disponibles libremente.
- **Acceso abierto:** Hacer que todos los resultados publicados sean de libre acceso para un uso e impacto máximos.

- **Open Notebooks:** Una práctica emergente que documenta y comparte el proceso experimental de prueba y error.

3.1. Datos abiertos

El mundo está presenciando una transformación global significativa, facilitada por la tecnología y los medios digitales, e impulsada por los datos y la información. Esta transformación tiene un enorme potencial para fomentar una investigación más transparente, responsable, eficiente, receptiva y eficaz. Solo una proporción muy pequeña de los datos originales se publica en revistas convencionales. A pesar de las políticas existentes sobre el archivo de datos, en la práctica actual, los datos se almacenan principalmente en archivos privados, no en repositorios institucionales seguros, y efectivamente se pierden para el público (y, a menudo, incluso para el investigador que generó los datos).

Esta falta de intercambio de datos es un obstáculo para la investigación por dos razones principales:

- Generalmente es difícil o imposible reproducir un estudio sin los datos originales.
- Los datos no pueden ser reutilizados o incorporados en un nuevo trabajo por otros investigadores si no pueden obtener acceso a ellos.

En consecuencia, existe una revolución global de datos en curso que busca promover la colaboración y la creación y expansión de programas de investigación efectivos y eficientes. Los datos abiertos están disponibles gratuitamente en Internet. Cualquier usuario puede descargar, copiar, analizar, reprocesar y reutilizar.

Entre los servidores para alojar los datos disponibles gratuitamente encontramos:

Zenodo: Repositorio de acceso abierto de propósito general desarrollado bajo el programa europeo OpenAIRE y operado por CERN

Academic Torrents: Permite compartir datos de investigación utilizando el protocolo BitTorrent

3.2 Entorno Computacional

Un entorno computacional es el sistema donde se ejecuta un programa. Esto incluye características de hardware (como la cantidad de núcleos en cualquier CPU) y características de software (como el sistema operativo, lenguajes de programación, paquetes de soporte, otras piezas de software instaladas, junto con sus versiones y configuraciones). Cada computadora tiene su entorno computacional, por esta razón las características en conjunto no coinciden en todos los casos y por ende no hay garantía de que el análisis pueda ejecutar o generar los mismos resultados, además los entornos computacionales evolucionan a medida que se actualizan los software.

Para que la investigación sea reproducible, el entorno computacional en el que se realizó debe capturarse de tal manera que otros puedan replicarlo. Una solución a esto son herramientas que los capturan y distribuyen de manera que otros puedan replicarlo. Hay varias formas de capturar entornos computacionales. Los principales son los sistemas de gestión de paquetes, el archivador, las máquinas virtuales y los contenedores. Se pueden dividir en dos categorías:

- Los que capturan solo el software y sus versiones utilizadas en un entorno (sistemas de gestión de paquetes).
- Las que replican un entorno computacional completo, incluido el sistema operativo y la configuración personalizada
 - **Máquinas Virtuales:** Es un software que simula un sistema de computación y puede ejecutar programas como si fuese una computadora real.
 - **Los Contenedores:** Ofrecen muchos de los mismos beneficios que las máquinas virtuales. Básicamente, actúan como máquinas completamente independientes que pueden contener sus propios archivos, software y configuraciones. La diferencia es que las máquinas virtuales incluyen un sistema operativo completo junto con todo el software asociado que normalmente se empaqueta con él, independientemente de si el proyecto utiliza este software asociado. Los contenedores sólo contienen el software y los archivos definidos

explícitamente dentro de ellos para ejecutar el proyecto que contienen. Esto los hace mucho más ligeros que las máquinas virtuales.

Para que poder asegurar el portal de reproducibilidad científica, es importante tener en cuenta el sitio en el servidor donde se va a almacenar toda la información y estructura (Hosting), también el dominio web para poder acceder, y finalmente una API que nos permita enlazar los procesos computacionales, importación de datos y publicaciones de los resultados científicos.

3.3 Cuadernos Interactivos

Una forma de computación interactiva es un entorno en el que los usuarios ejecutan código, ven lo que sucede, modifican y repiten. Son una de las herramientas más interactivas, sencillas y fáciles de reproducir. Jupyter Notebook es un ejemplo de estos cuadernos y es de código abierto en el cual se usa para combinar código de software, resultados computacionales, texto explicativo y recursos multimedia en un solo documento.

Los cuadernos abiertos tienen el beneficio adicional de aumentar la calidad de los resultados científicos al obligar a los investigadores a ser cuidadosos, minuciosos y explícitos. Hacer que la investigación sea abierta tiene el beneficio adicional de aumentar la probabilidad de que los errores cometidos en una investigación se detecten rápidamente, en lugar de en el futuro. Las soluciones inmediatas tendrán un impacto mucho menor en un proyecto de investigación, lo que ahorrará tiempo de investigación, dinero de laboratorio y orgullo.

3.4 Control de Versiones

El control de versiones nos ayuda a comprender qué cambios hicimos en el pasado o por qué hicimos un análisis específico de la forma en que lo hicimos, incluso semanas o meses después. Con la ayuda de comentarios y mensajes de confirmación, cada versión puede explicar qué cambios contiene en comparación con las versiones anteriores. Esto es útil cuando compartimos nuestro análisis (no solo datos) y lo hacemos auditable o reproducible, lo cual es una buena práctica científica.

Finalmente, el control de versiones es invaluable para proyectos colaborativos donde diferentes personas trabajan en el mismo código simultáneamente y se basan en el trabajo de los demás. Permite realizar un seguimiento de los cambios

realizados por diferentes personas y puede combinar automáticamente el trabajo de las personas al tiempo que se ahorra una gran cantidad de esfuerzo para hacerlo manualmente. Además, los servicios de hospedaje de control de versiones, como GitHub, brindan una forma de comunicarse y colaborar de una manera más estructurada, como en solicitudes de extracción, revisiones de código y problemas.

Todas estas herramientas que se mencionaron anteriormente fomentan la reproducibilidad computacional al simplificar la reutilización del código.

4.- Beneficios de la Investigación Abierta

- Las prácticas abiertas también benefician a los investigadores que las propagan.
- Las prácticas abiertas pueden facilitar la conexión de los investigadores al aumentar la capacidad de descubrimiento y visibilidad del trabajo de uno, facilitando el acceso rápido a nuevos datos y recursos de software, y creando nuevas oportunidades para interactuar y contribuir a proyectos comunes en curso.

5.- Presupuesto

Se presenta un presupuesto para desarrollar una infraestructura sostenible y reproducible para un grupo pequeño de Investigación durante el primer año, tratando de integrar todos los criterios necesarios para la investigación científica y su reproducibilidad. Los criterios necesarios van desde el alojamiento de los datos hasta la publicación de los resultados finales. Estos se organizan según la tabla 1.

Tabla 1: Criterios necesarios para la investigación científica y su reproducibilidad.

Número de usuarios potenciales	~ 1000
Número de posibles creadores de contenido	~1 - 4
Privacidad de datos	Bajo a medio: cualquiera puede acceder a los conjuntos de datos, pero el análisis en curso está sujeto a un

	proceso de embargo durante un período de tiempo finito. Pueden utilizar datos abiertos de fuentes externas. Los recursos finales deben ser de acceso abierto
Características de los datos	<ul style="list-style-type: none"> • ~ Homogéneo • imágenes grandes (~ 50%) • Tablas CSV (~ 50%)
Tamaños de datos	~10 TB

Se tiene un presupuesto de 100 monedas para un año y se distribuye de la siguiente manera:

¿Es mejor comprar o alquilar el Hardware?

Es una decisión difícil, debido a que si se renta el Hardware a largo plazo sale más costoso que comprarlo, además el modelo de alquiler de tecnología tienes sus ventajas, ya que reducen la inversión necesaria para poder utilizar un equipamiento determinado, ofrece una mayor flexibilidad ante los cambios que se produzcan en las necesidades del usuario, y permite dejar de alquilar un dispositivo o producto concreto para pasar a alquilar la versión superior o más actual tan pronto como esté a la venta. Sin embargo tiene sus desventajas, pero para nuestro proyecto teniendo en cuenta que es para un año es mejor comprarlo.

40(monedas) Hardware

20(monedas) Entornos computacionales

20(monedas) Internet de Ultra Alta velocidad

18(monedas) Hosting y dominios

2(monedas) Gastos de Luz tomando en cuenta el servicio en Venezuela

6.- Conclusión

Se garantizo la reproducibilidad y sostenibilidad del proyecto porque se adoptó y se amplió los estándares abiertos existentes, se pudo compartir los datos con el fin de fomentar la innovación abierta por parte del grupo de investigación, se usó plataformas abiertas existentes, a fin de ayudar a automatizar el intercambio de datos, conectar herramienta o sistema con otros y agregar flexibilidad para adaptarse a necesidades futuras, se desarrolló el nuevo código de software para que sea de fuente abierta, que cualquiera pueda ver, copiar, modificar y compartir, y distribuir el código en repositorios públicos y se posibilitó la innovación al compartir

libremente sin restricciones, colaborando de manera amplia y co-creando herramientas.

Referencias

- [1] The Turing Way Community, “Overview of Reproducible Research”. 2020. Website:
<https://the-turing-way.netlify.app/reproducible-research/overview.html>
[July 28]

- [2] The Turing Way Community, “Open Data”. 2020, Website:
<https://the-turing-way.netlify.app/reproducible-research/open/open-data.html>. [July 30]

- [3] The Turing Way Community, “Reproducible Environments”, 2020
<https://the-turing-way.netlify.app/reproducible-research/renv.html>. [July 31 25]