Vol. 19, n.° 4, pp.315-324, 2020
Revista UIS Ingenierías
Página de la revista: revistas.uis.edu.co/index.php/revistauisingenierias

Editorial

Academia, datos y reproducibilidad de la ciencia
Academy, data and science reproducibility

Alexander Martínez-Méndez [1], Luis Alberto Nuñez [2]

[1] Escuela de Ingeniería de Sistemas e Informática, Universidad Industrial de Santander, Bucaramanga 680002, Colombia. Orcid: 0000-0002-1559-9015. Correo electrónico: alexanderm2@protonmail.com
2Escuela de Física, Universidad Industrial de Santander Bucaramanga 680002, Colombia y
Departamento de Física, Universidad de Los Andes, Mérida 5101, Venezuela.
Orcid: 0000-0003-4575-5899. Correo electrónico: lnunez@uis.edu.co

Publicado en línea: 11 septiembre, 2020.

Resumen

La pandemia de COVID-19 ha catalizado una serie de prácticas académicas y cuando la superemos, la geografía y, sobre todo la dinámica de los grupos de investigación será otra. Quizá seremos más eficaces, pero el rasgo que será más indeleble de ese cambio lo constituirá la incorporación cotidiana de prácticas de ciencia abierta, repetible y reproducible. El cambio cualitativo de la investigación a la "e-investigación", describe nuevas formas de producción y diseminación del conocimiento e impone nuevas metodologías para manejar, administrar, analizar y preservar este "diluvio de datos". La repetibilidad y la reproducibilidad de los resultados científicos en la ciencia centrada en datos ha sido detectado como un un problema creciente. Las publicaciones académicas están llamadas a transformarse, exigiendo acceso a los datos y aplicaciones computacionales que soportan los resultados. En esta nota editorial comentamos y reflexionamos sobre algunas de las metodologías y herramientas que se disponen para promover la repetibilidad/reproducibilidad de los experimentos. Discutimos el enfoque metodológico que emerge del Instituto Turing del Reino Unido. Presentamos una evaluación de las herramientas para construir repositorios de datos que existen actualmente.

Palabras clave: reproducibilidad; ciencia abierta; repositorios de datos; acceso abierto.

Abstract

The COVID-19 pandemic has catalyzed several academic practices, and when it ends, the geography and especially the dynamics of research groups will be completely different. We will have other practices, and, in many cases, we will be more effective. Perhaps the most indelible feature of that change will be the day-to-day incorporation of open, repeatable, and replicable science practices. The qualitative change from research to "e-research" describes a new way to produce and disseminate knowledge. It also imposes new methodologies to manage, administer, analyze and preserve this "data deluge". The repeatability and reproducibility of scientific results in data-centred science become a growing and significant issue. Undoubtedly, academic publications should be transformed, demanding open access to data and direct availability to computer codes and applications that support the results. In this editorial, we comment and reflect on some of the methodologies and tools available to promote the repeatability/reproducibility of experiments.

Keywords: reproducibility; open science; data repositories; open access.

ISSN impreso: 1657 - 4583. ISSN en línea: 2145 – 8456, **CC BY-ND 4.0**
Como citar: A. Martínez-Méndez, L. Núñez, "Academia, datos y reproducibilidad de la ciencia," *Rev. UIS Ing.*, vol. 19, no. 4, pp. 315-324, 2020, doi: https://doi.org/10.18273/revuin.v19n4-2020026

1. Introducción

La pandemia de COVID-19 ha catalizado una serie de prácticas en la academia. Aprendimos que podemos manejar, medianamente y hasta bien, gran parte de las actividades docentes. Nos convencimos que las ideas pueden fluir sin necesidad de estar impulsadas por la movilidad y presencialidad en congresos y/o reuniones científicas. Poco a poco vamos comprendiendo que muchos de esos aprendizajes llegaron para quedarse. Cuando la pandemia pase, la geografía y, sobre todo la dinámica de los grupos de investigación será otra. Tendremos otras prácticas y, en muchos casos, seremos más eficaces.

Quizá el rasgo que será más indeleble de ese cambio lo constituirá la incorporación cotidiana de prácticas de ciencia abierta, repetible y reproducible. El genoma del virus SARS-CoV-2 se publicó libremente a una velocidad sin precedentes[1]. Los grupos de investigación de todo el mundo modelan la pandemia basándose en datos publicados por la Universidad John Hopkins[2]. La movilidad mundial, registrada en nuestros teléfonos inteligentes, es pública está disponible[3].

Los repositorios institucionales [1] y temáticos se convierten en los centros de diseminación y socialización de las experiencias/resultados de los distintos grupos de investigación. La biblioteca pública digital sigue reconquistando los espacios que habían sido ocupado por las editoriales comerciales [2]. Las comunidades de Física de Altas Energías, Matemáticas y Astrofísica utilizan los repositorios temáticos cotidianamente [3]. Pero la pandemia impuso su uso en biomedicina, que habían sido reticente a incorporarse a esta práctica de difundir libremente las ideas antes de que aparezcan publicadas en revistas especializadas. Dos repositorios temáticos de biomedicina registran un uso explosivo[4] y seguro esa práctica será permanente.

Esta transformación indetenible viene desde la década de los 70's cuando cambió el modo de producción del sistema capitalista y pasamos de una economía industrial a una informacional [4, 5]. La información transformó la economía en el mismo sentido que la industria modificó la actividad económica en industrial. La academia, la actividad científica y tecnológica también se han venido transformado en términos metodológicos, funcionales y, sobre todo, por la manera como nos hemos venido organizando para crear y diseminar el conocimiento que producimos [6, 7].

El cambio cualitativo de la investigación a la "*e*-investigación", describe esas nuevas formas de producción y diseminación del conocimiento e impone nuevas metodologías para manejar, administrar, analizar y preservar este "diluvio de datos" (ver [8, 9, 10] y las referencias allí citadas).

Uno de los elementos claves en esta nueva era de ciencia centrada en datos es la repetibilidad y reproducibilidad de los hallazgos. Es indispensable que un investigador pueda encontrar las publicaciones de su interés y sea redirigido automáticamente a los datos (medidos y/o simulados) que la generaron. Ese investigador también deberá poder acceder a las aplicaciones computacionales que produjeron esos resultados. Los nuevos datos generados a partir del análisis anterior (y el nuevo documento que reporte los resultados) se deben poder preservar para poder ser fácilmente encontrados. Esta búsqueda de datos y la generación de resultados que se convierten otros datos que serán utilizados por otros investigadores, genera un círculo virtuoso: ...datos → información → conocimiento → datos... y permite iniciar nuevamente el ciclo [11].

La repetibilidad y la reproducibilidad de los resultados en la ciencia centrada en datos han sido detectados como uno de los problemas crecientes en la actividad científica. Sin duda, las publicaciones académicas están llamadas a transformarse, exigiendo acceso a los datos y aplicaciones computacionales que soportan los resultados. Las revistas más prestigiosas han alertado sobre esta crisis [12, 13] y la comunidad lo reconfirma [14][Ver figura 1]. Esta crisis motiva retiros de artículos aparecidos en prestigiosas publicaciones y algunas renombradas asociaciones científicas comienzan a generar un código de conducta para sus afiliados [15]. El retiro de artículos ha aumentado al exigir la presentación de los datos y/o códigos que generaron los resultados. Los retiros por falta de datos, códigos, plagio y otros tipos de fraudes o conductas indebidas comienzan a ser reportados públicamente, implementando un mecanismo de autocontrol de la comunidad científica[5].

En esta nota editorial queremos comentar algunas de las metodologías y herramientas que se disponen para promover la repetibilidad/reproducibilidad de los experimentos. En la próxima sección discutiremos *The*

[1] https://www.gisaid.org/

[2] https://coronavirus.jhu.edu/data

[3] Los datos de dispositivos Apple-IOS pueden ser accedidos en https://www.apple.com/covid19/mobility mientras que Google registra los de Android en https://www.google.com/ covid19/mobility?hl=en

[4] https://www.medrxiv.org y https://www.biorxiv.org

[5] En este sitio web se reportan gran parte de esos artículos fraudulentos http://retractionwatch.com

Touring Way, una guía para la producción de conocimiento reproducible, ético, colaborativo e inclusivo. En la sección 3 haremos un listado de las alternativas para generar repositorios de datos que apoyen a las publicaciones científicas. Seguidamente, en la sección 4, describiremos las publicaciones ejecutables, como una tendencia para facilitar la reproducibilidad de las simulaciones y análisis de datos. En la sección 5 haremos una breve descripción de los impactos de la ciencia abierta en la reproducibilidad y replicabilidad de la actividad científica. Finalmente, en la sección 6 hacemos una reflexión general de alguno de los impactos de los cambios que estamos experimentando y un recuento de los principales puntos desarrollados.

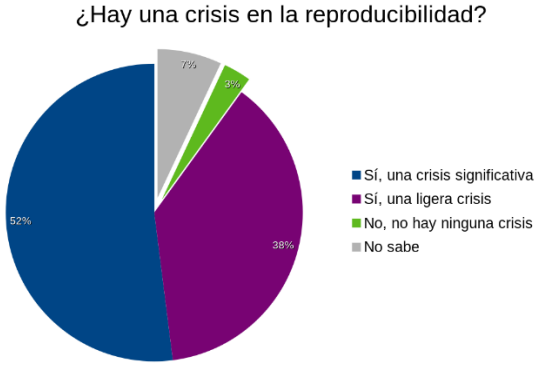

Figura 1. Resultados a la pregunta ¿Hay una crisis en la reproducibilidad? en encuesta realizada por la revista Nature. 1,500 scientists lift the lid on reproducibility,[14].

Impresiona el hecho que solo el 10% de los encuestados considera que no ha crisis o no sabe si existe.

2. *The Touring Way*

The Touring Way es una guía metodológica –desde y hacia la comunidad científica– para la producción de conocimiento reproducible, ético, colaborativo e inclusivo [16].

En ese documento se describen con mucho detalle los tres elementos esenciales para la reproducibilidad de los experimentos científicos:

• Los datos generados o producidos en el marco de la investigación;

• Los códigos computacionales o el software utilizado para el análisis e interpretación de los datos;

• La información sobre las características y estado de las herramientas, equipos y/o dispositivos usados en la investigación. En la figura 2 esquematizamos estos tres elementos. La definición de reproducibilidad en la literatura toma distintos matices pero, en esencia, todos son muy similares. Para este documento supondremos las definiciones establecidas desde *The Turing Way*. En este sentido, dependiendo de los datos y análisis utilizados, tendremos cuatro escenarios para la reproducibilidad de un experimento[6].

Figura 2. La reproducibilidad de un experimento está determinada por el acceso y posibilidad de uso de los conjuntos de datos, los códigos computacionales. También, es indispensable una completa descripción del escenario y ambiente dónde se desarrollaron tanto el experimento como los análisis de datos [16].

[6] Estamos considerando como protocolo científico aquel conjunto de reglas que generan las mediciones (montaje experimental, condiciones de registro de los instrumentos) o las suposiciones o propuestas que respaldan a determinados sistemas o códigos computacionales que generan los datos simulados.

En la figura 3 reproducimos la ilustración de la matriz de reproducibilidad, donde cruzamos la importancia del análisis y los datos sobre los cuales éste opera [16].

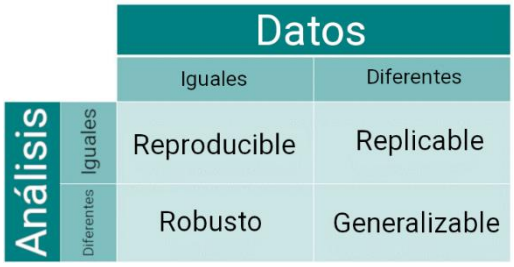

Figura 3. Definición de reproducibilidad por *The Turing Way* basada en la disponibilidad o no de los conjuntos de datos y análisis de un experimento. Con mayor frecuencia, un experimento será generalizable si se puede realizar sobre diferentes conjuntos de datos y con diferentes metodologías de análisis [16].

En el primer escenario, un experimento será reproducible si contamos tanto con los mismos datos como con los mismos tipos de análisis. En el siguiente escenario, el experimento será replicable si los análisis se pueden llevar a cabo sobre diferentes conjuntos de datos. Un experimento será robusto si diferentes análisis llegan a las mismas conclusiones, operando sobre el mismo conjunto de datos. Finalmente, un experimento será generalizable si se puede realizar sobre diferentes conjuntos de datos y con diferentes metodologías de análisis.

2.1. Datos de investigación

Actualmente, las infraestructuras de cómputo y una variedad de instrumentos conectados a la red generan enormes cantidades de datos. Este alud de mediciones convierte a los instrumentos en herramientas informáticas y la experimentación en minería de datos. Esto genera un cambio metodológico en la generación de conocimiento científico, cada vez más centrado en grandes volúmenes de datos.

La comunidad científica ha reconocido la importancia de gestionar adecuadamente este océano de datos y, por consiguiente, la necesidad de abordar los desafíos propuestos en estos escenarios. Las soluciones desarrolladas han girado en torno al descubrimiento, acceso, análisis y preservación de los datos medidos y/o simulados. Simultáneamente, se han desarrollado/adoptado diversos estándares para facilitar la interoperabilidad de las herramientas (hardware y software) informáticas. Múltiples iniciativas buscan definir caminos exitosos para la gestión de datos. En la figura 4 ilustramos los aspectos más resaltantes descritos en *The Turing Way* y en otras importantes referencias [16, 17]. Adicionalmente los principios FAIR (*Findable, Accesible, Interoperable, Reusable*) [18] proveen una ruta muy clara para avanzar en la gestión de datos científicos.

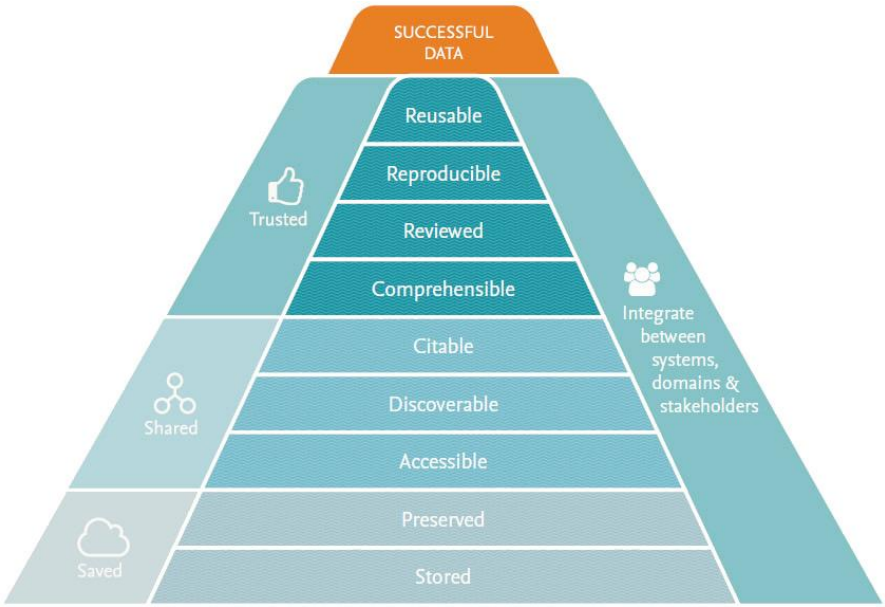

Figura 4. Aspectos a tener en cuenta para hacer una exitosa gestión de datos de investigación. En esta figura se representa la "ruta" desde su adquisición hasta su preservación, teniendo como punto de referencia la integración interoperabilidad entre los distintos sistemas [17].

Una estrategia de gestión de datos debe partir siempre del conocimiento de las características de los conjuntos de datos. Los *Planes de Gestión de Datos*, PGD, son las herramientas metodológicas adecuadas para lograr este conocimiento. En un PGD se responde a los interrogantes ¿Qué datos generamos y/o usaremos en nuestro proyecto? y ¿Cómo gestionaremos estos conjuntos de datos? Así, un PGD debe ser el punto de partida para cualquier experimento científico y su estrategia de gestión de datos. Los repositorios de datos de investigación son una herramienta de gran valor en los PGD y contribuyen a la gestión, preservación y diseminación de datos producto de la investigación. De estos sistemas nos ocuparemos en el siguiente capítulo.

2.2. Códigos computacionales

La avalancha de registros de todo tipo viene generada por experimentos de escala mundial – aceleradores de partículas, red de observatorios terrestres y satelitales e infinidad de los más variados sensores, IoT, dispersos geográficamente, pero interconectados–, los cuales desbordan toda capacidad de manejo que no sea mediante uso intensivo de sofisticados sistemas computacionales y de comunicación.

Es decir, para estudiar un fenómeno en estos escenarios no es suficiente contar con datos de buena calidad, ya que el simple hecho de observarlos nos tomaría toda una vida. Es imperioso el uso de las Tecnologías de Información y Comunicación, de sistemas computacionales que permitan el análisis y nos auxilien en su interpretación. En estos códigos computacionales se plasman y automatizan un conjunto de instrucciones que buscan extraer, de forma automática, información, relaciones, tendencias y/o eventos sumergidos ese mar de datos.

Entonces surge la dificultad de la reproducibilidad computacional. Esto es el poder tener acceso a los códigos computacionales que descubrieron esas tendencias y relaciones sumergidas. Más allá de los diversos lenguajes de programación para la creación de los códigos computacionales –y de los distintos paradigmas en los que se enmarcan– resaltamos acá los sistemas de control de versiones (CVS, por sus siglas en inglés) en la creación y gestión de códigos computacionales. Los CVS nacieron con el desarrollo de software y nos permiten acceder a sus últimas versiones, haciendo una trazabilidad de los cambios realizados por distintos autores. Actualmente estos sistemas han evolucionado y permiten un trabajo colaborativo mucho más eficiente.

2.3. Entornos de trabajo

La *e*-investigación impone un manejo automático de grandes volúmenes de datos. Para ser descubiertos, accedidos y analizados, los datos deben ser fácilmente identificables y, sobre todo ser confiables. Una adecuada documentación sobre el muestreo, procedimientos analíticos, anomalías y calidad de los datos ayudará a que esos datos puedan ser correctamente interpretados y reinterpretados, en el futuro. La interacción automática entre repositorios de datos y sistemas computacionales serán cotidianas en el futuro inmediato, por lo tanto, se requiere una documentación de los datos que sea reconocible e interoprable por variados sistemas computacionales.

Por ello el entorno donde se realizan los experimentos científicos es el tercer elemento determinante para su reproducibilidad. Los parámetros y configuraciones de equipos y/o dispositivos usados, así como también información sobre el flujo de trabajo, permitirán o no la reproducibilidad del experimento. Por lo tanto, es importante capturar y preservar desde los apuntes en cuadernos de laboratorio hasta la temperatura del ambiente.

En cuanto a entornos computacionales, idealmente, la mejor forma de facilitar la reproducibilidad es ofrecer el entorno como servicio. Sin embargo, un servicio de éstos representa una gran inversión para un alto porcentaje de (pequeños y medianos) centros de investigación. Una opción más viable es empaquetar las herramientas de software con tecnologías de contenerización o virtualización.

3. Repositorios para gestión de datos

Muchos de los datos registrados por los grupos de investigación nunca aparecerán publicados. Cuando finalizan los experimentos, muchos de esos datos se perderán o serán enviados a reservorios nacionales (o internacionales) que nada tuvieron que ver con su producción. Más aún, muchas de las decisiones y criterios que generaron su producción y depuración quedarán escritos en una inmensa correspondencia electrónica que nadie podrá consultar [19].

Es imperioso planificar y construir repositorios de datos que los almacenen mientras se produzcan, que conserven la traza de las decisiones y criterios que los generaron y que los preserven en el tiempo [19, 20].

Adicionalmente, es indispensable utilizar ambientes de mensajería instantánea que permita recuperar los diálogos, mensajes, datos y publicaciones intercambiados por los investigadores que estarán, cada vez más distribuidos en diversas partes del mundo. El correo electrónico queda relegado por estos sistemas de mensajería utilizados cada vez más frecuentemente por los grupos de investigación. Los sistemas de repositorios nos permiten almacenar, catalogar, preservar y compartir conjuntos de datos y son la herramienta indispensable para la gestión de datos en entornos de ciencia abierta. La elección de una estrategia de gestión de datos, como lo es el sistema de repositorio más adecuado para su contexto, estará varios escalones arriba en la pirámide propuesta en figura 4.

A continuación, presentamos una evaluación comparativa de cuatro de los sistemas de repositorio más utilizados en la actualidad, *Zenodo, Dataverse, Dspace y Ckan.* Acá buscamos elegir el mejor software para repositorios de datos en pequeños o medianos grupos de investigación.

Este estudio se basó en las enseñanzas y comentarios entorno a la gestión de datos de investigación encontradas en los trabajos [21, 22, 23, 24, 25, 18], especialmente de [17, 16].

La evaluación se realizó en base a parámetros funcionales y no funcionales de los softwares, entre los que destacan:

- La capacidad de integración a otros sistemas mediante API's y estándares de interoperabilidad.
- El uso de identificadores persistentes y la inclusión de metadatos que permitan búsqueda, difusión y autoría de los conjuntos de datos.
- La documentación disponible para su uso y administración.
- Finalmente, parámetros propios del desarrollo como su licencia, la comunidad detrás del desarrollo y el lenguaje de programación utilizado.

CKAN[7], (*Comprehensive Knowledge Archive Network*), es una solución de administración de datos y un portal de código abierto con diversas funcionalidades. Proporciona una forma simplificada de hacer que los contenidos sean visibles y presentables, esto junto a un rico conjunto de metadatos para aumentar su valor y facilitar su descubrimiento. En esta herramienta se destaca la visualización de diversos tipos de contenidos en su versión web, la facilidad de implementar plugins para aumentar sus funcionalidades y finalmente, el API más completo de las herramientas evaluadas.

Dataverse[8], es una aplicación web de código abierto para compartir, preservar, citar, explorar y analizar datos de investigación [26]. Desarrollada por IQSS (*Institute for Quantitative Social Science*) en la Universidad de Harvard y con el apoyo de un amplio número de desarrolladores a nivel mundial [27].

Zenodo[9], es un servicio ofrecido por el CERN (por sus siglas en inglés del Centro Europeo de Investigaciones Nucleares) en el marco de la política de datos abiertos del proyecto OpenAIRE[10]. *Zenodo* busca potenciar las capacidades de bibliotecas digitales, garantizando accesibilidad, interoperabilidad y usabilidad en los contenidos publicados [28]. Es relevante mencionar que *Zenodo* es la capa de la interfaz web, éste en el fondo utiliza el software *Invenio*[11], software escrito con el *framework* de desarrollo web Flask [29]. Se destaca de esta herramienta su gratuidad, la integración con el servicio GitHub y la asignación de identificadores digitales persistentes.

DSpace[12], es un software de código abierto para preservar contenidos digitales [30]. Esta herramienta ha sido desarrollada en sus inicios por el *Massachusetts Institute of Technology* y los laboratorios de *Hewlett Packard Company*. En la actualidad voluntarios contribuyen en su desarrollo y mantenimiento. Es uno de los sistemas de repositorios institucionales más usado según la base de datos OpenDOAR [31] y su principal utilización se centra en la creación de repositorio digitales para bibliotecas. Se destaca negativamente para este sistema la lentitud en la generación de nuevas versiones.

4. Publicaciones ejecutables

Una estrategia para presentar de manera transparente el análisis de los datos son las publicaciones ejecutables e interactivas. La idea es ofrecer una interfaz para publicaciones como piezas de software dinámica, que combinen texto, datos (en bruto o procesados) y el código utilizado para el análisis. De esta forma el lector pueda interactuar con el texto, las gráficas explorando casos límites y otras posibles conclusiones a partir de los mismos conjuntos de datos y aplicaciones computacionales [32].

[7] https://ckan.org/
[8] https://dataverse.org/
[9] https://zenodo.org/
[10] https://www.openaire.eu/
[11] https://invenio-software.org/
[12] https://duraspace.org/DSpace/

La idea de los artículos ejecutables que permitan compartir los códigos para el análisis de datos ha estado presente en la comunidad editorial desde hace al menos una década. Fue motivo de un concurso mundial promovido por Elsevier, una de las mayores editoriales científicas del mundo [13]

Actualmente, muchas revistas alientan a los autores a compartir los datos, pero sólo para una minoría es obligatorio y rara vez proveen una metodología sobre la forma como deben compartirse [33]. Como hemos dicho, la reproducibilidad computacional requiere del acceso a los códigos y sistemas informáticos que permiten el análisis de los datos y, en este terreno, la adopción por las revistas de esta metodología de validación de resultados está aún menos desarrollada.

La falta de ambientes y herramientas estandarizadas para compartir los componentes de la investigación, la presión por "publicar o perecer" genera mayores obstáculos que dificultan la reproducibilidad. Existen algunos ambientes que, sin estar integrados ni pensado para estos fines pueden representar un paliativo temporal. Una herramienta muy conocida en el campo de la investigación reproducible es la interfaz de "cuadernos" Jupyter (por Julia, Python y R) [34]. En este ambiente los lectores pueden abrir fácilmente documentos reproducibles y manipular el código para ver los resultados cambiantes. La documentación en el lenguaje de etiquetas *Markdown* combina el código con textos elaborados interpretados en LATEX. Los documentos pueden ser intercambiados, renderizados a HTML que ejecuta el código y muestra una documentación muy rica. Cuando se combina con con sistemas CVS tipo *GitHub* se puede hacer la trazabilidad de cambios y obtener las últimas versiones. Tal y como se muestra en la referencia [32] el binomio Jupyter-Git puede ser un buen comienzo para que las revistas científicas promuevan la reproducibilidad de los artículos que publican.

5. Ciencia abierta

La ciencia abierta es una visión de la investigación científica en la cual sus productos –i.e publicaciones, datos y códigos computacionales–, y su difusión son accesibles a la sociedad en todos sus niveles. La ciencia abierta es el conocimiento transparente, accesible y colaborativo. Desde al acceso abierto al trabajo colaborativo con la sociedad, la *Ciencia Abierta* nos plantea un escenario prometedor indispensable para la reproducibilidad.

"*Queremos más ciencia y la queremos abierta*" exclama la *Declaración de Panamá* sobre *Ciencia Abierta* y refuerza esa perspectiva de la ciencia como motor fundamental de la sociedad, proveyendo la visión amplia y latinoamericana. Más allá de dar una definición de la Ciencia Abierta, la declaración de Panamá hace precisiones sobre los elementos que la componen y los ilustramos en la figura 5.

La ciencia abierta incluye el acceso abierto a las publicaciones, a los datos, sumando esfuerzos con el movimiento de código abierto y software libre. Claramente, hoy la ciencia requiere de una creciente infraestructura para procesar ingentes cantidades de datos, a partir de los cuales debemos identificar relaciones y dependencias. Por ello, el acceso a los códigos es indispensable para reproducibilidad de los resultados científicos.

Figura 5. Ejes fundamentales de la ciencia abierta. Tomado de *Declaración de Panamá* sobre Ciencia Abierta [35]. La ciencia abierta incluye el acceso abierto a las publicaciones, a los datos, sumando esfuerzos con el movimiento de código abierto y software libre.

6. Conclusiones

Quizá no tengamos una conciencia clara de los profundos cambios que ha catalizado la pandemia covid-19. Estos cambios venían dándose lentamente, pero la pandemia los aceleró. Es tal la cantidad de información a la cual hoy tienen acceso nuestros estudiantes, que debemos repensar las metodologías que utilizamos cotidianamente en su formación. Nuestra función como docentes habrá de focalizarse en la enseñanza de los principios básicos en ciencias y humanidades, proveyendo la capacitación necesaria para que puedan encontrar en la red la información pertinente y valorar su calidad [36].

[13] Puede consultar los premios en: https://www.elsevier.com/about/press-releases/science-and-technology/elsevier-announces-winners-of-the-executable-paper-grand-challenge

La confianza en la veracidad de los datos está relacionada con su curaduría y con la calidad de los metadatos que muestren. Cada vez más, será indispensable adhesión a protocolos de registro de datos durante la investigación. Se confía en un investigador o un grupo de investigación basado en la meticulosidad y seguimiento estricto de estos protocolos de descubrimiento con los cuales respalda o respaldan sus resultados [37]. La confianza en la consulta automática entre repositorios de datos y entre aplicaciones y repositorio se fundamentará en la calidad de los metadatos consecuencia del seguimiento de los protocolos.

La reproducibilidad de experimentos científicos es la base de la ciencia misma. Ofrecer herramientas que faciliten la reproducibilidad o replicabilidad de un experimento debe ser un objetivo a cumplir. Como se expresa en *The Turing Way*, la reproducibilidad "debe ser tan fácil que lo difícil sea no hacerlo".

Existe una variedad de posibles soluciones para la implementación y utilización de repositorios de datos, la elección de esta herramienta deberá hacerse a partir de un profundo análisis del experimento y su contexto. La herramienta con las mejores características observadas en este trabajo de investigación ha sido *CKAN*. Sin embargo, consideramos que *Zenodo* es una alternativa interesante y sería la elección correcta en escenarios donde no se cuente con los recursos necesarios para desplegar y administrar un servicio de repositorio de datos.

Estamos ante un nuevo escenario en la producción de conocimiento y se nota con gran fuerza en los requisitos para acceder a fondos de financiación. En la gran mayoría de convocatorias el acceso abierto al conocimiento y a los productos generados en una investigación se ha vuelto más que frecuente. Tenemos que reforzar la idea que los datos generados con financiamientos públicos son patrimonio de la humanidad y deben estar accesibles y disponibles tan amplia y directamente como se pueda. Esta visión contrasta con la actitud de investigadores y grupos de investigación que consideran los datos como su patrimonio y, sobre todo, se enfrenta a la reciente posición de muchos editores, quienes comienzan a exigir los datos que respaldan las publicaciones, haciéndoles extensivo el derecho de reproducción (*copyright*), con la consecuente restricción para su reutilización. Quizá el acceso a los datos pueda ser limitado si su utilización arriesga la seguridad de individuos o especies, compromete derechos de confidencialidad, o viola prerrogativas para su explotación temporal por quienes los recolectaron o generaron [38].

Referencias

[1] J. A. Dávila, L. Núñez, B. Sandia, R. Torréns, "Los repositorios institucionales y la preservación del patrimonio intelectual académico," *Interciencia*, vol. 31, no. 1, pp. 22–29, 2006.

[2] L. Núñez, "La reconquista digital de la biblioteca pública," *Interciencia*, vol. 27, no. 4, pp. 195– 201, 2002.

[3] P. Ginsparg, "Arxiv at 20," *Nature*, vol. 476, no. 7359, pp. 145–147, 2011.

[4] M. Castells, *The Rise of the Network Society*. Cambridge, MA, USA: Blackwell Publishers, Inc., 2000.

[5] M. Castells, *The Internet Galaxy*. Oxford UK: Oxford University Press, 2001, doi: 10.1093/acprof:oso/9780199255771.001.0001

[6] M. Nielsen, *Reinventing Discovery: The New Era of Networked Science.* Princeton University Press, Oct. 2011.

[7] L. A. Núñez, "Ciencia abierta y de datos: retos y realidades," *Deslinde*, vol. 58, pp. 69–75, 2015.

[8] T. Hey, A. E. Trefethen, "e-science and its implications," *Phil. Trans. R. Soc. Lond. A*, vol. 361, pp. 1809–1825, 2003.

[9] I. Foster, "Service-oriented science," *Science*, vol. 308, pp. 814–817, May 2005.

[10] T. Hey, A. E. Trefethen, "Cyberinfrastructure for e-science," *Science*, vol. 308, pp. 817–821, May 2005.

[11] R. Barbera, B. Becker, C. Carrubba, G. Inserra, S. Jalife-Villalón, C. Kanellopoulos, K. Koumantaros, R. Mayo-García, L. Núñez, O. Prnjat, R. Ricceri, M. Rodriguez-Pascual, A. Rubio Montero, F. Ruggieri, "Chain-reds dart challenge," in *IV Conferência Internacional sobre Bibliotecas e Repositórios Digitais (BIREDIAL) y IX Simpósio Internacional de Bibliotecas Digitais (SIBD)(Porto Alegre, Brasil, 2014)*, 2014.

[12] M. Baker, "Reproducibility crisis," *Nature*, vol. 533, no. 26, pp. 353–66, 2016.

[13] J. Berg, "Progress on reproducibility," *Science*, vol. 359, no. 6371, pp. 9–9, 2018, doi: 10.1126/science.aar8654

[14] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, no. 7604, pp. 452–454, May 2016, doi:10.1038/533452a

[15] National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science.* Washington, DC: The National Academies Press, 2019, doi: 10.17226/25303

[16] B. Arnold, L. Bowler, S. Gibson, P. Herterich, R. Higman, A. Krystalli, A. Morley, M. O'Reilly, K. Whitaker, Community The Turing Way, *The Turing Way: A Handbook for Reproducible Data Science.* Zenodo, Mar. 2019. [Online]. Available: https://zenodo.org/record/3233986

[17] A. de Waard, H. Cousijn, I. J. Aalbersberg, "10 aspects of highly effective research data," *Elsevier Connect,* Dec. 2015. [Online]. Available: https://www.elsevier.com/connect/ 10-aspects-of-highly-effective-research-data

[18] M. D. Wilkinson, *et al.,* "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160018, Dec. 2016, doi: 10.1038/sdata.2016.18

[19] J. Gray, A. Szalay, "The world-wide telescope," *Commun. ACM*, vol. 45, no. 11, pp. 50–55, 2002.

[20] H. Karasti, K. Baker, E. Halkola, "Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the long term ecological research (lter) network," *Computer Supported Cooperative Work (CSCW)*, vol. 15, no. 4, pp. 321–358, August 2006, doi: 10.1007/s10606-006-9023-2

[21] W. Haak, "4 principles for unlocking the full potential of research data," *Elsevier Connect*, Mar. 2019. [Online]. Available: https://www.elsevier.com/connect/ 4-principles-for-unlocking-the-full-potential-of-research-data# contributors

[22] R. Delevante, "5 trends in research data management," *Elsevier Connect*, Sep. 2019. [Online]. Available: https://www.elsevier.com/connect/ 5-trends-in-research-data-management

[23] L. Willems, "6 insights from leading universities on managing research data effectively," *Elsevier Connect*, Apr. 2019. [Online]. Available: https://www.elsevier.com/connect/ 6-insights-from-leading-universities-on-managing-research-data-effectively

[24] M. van der Graaf, L. Waaijers, "A Surfboard for Riding the Wave. Towards a four country action programme on research data. A Knowledge Exchange Report," *Tech. Rep*., Nov. 2011. [Online]. Available: https://libereurope.eu/ a-surfboard-for-riding-the-wave-towards-a-four-country-action-programme-on-research-data/

[25] B. Centre for Science and Technology Studies (CWTS), "Open data: The researcher perspective," *Tech. Rep.,* Apr. 2017. [Online]. Available: https://www.elsevier.com/__data/assets/pdf_file/0004/281920/Open-data-report.pdf

[26] M. Crosas, "The dataverse network: an opensource application for sharing, discovering and preserving data," *D-lib Magazine*, vol. 17, no. 1, p. 2, 2011

[27] Dataverse, "Sitio oficial," 2018. [Online]. Available: https://dataverse.org/

[28] K. Nowak, "Zenodo - research. shared. second part of the open research data in h2020 & zenodo repository"webinar." Oct 2016.

[29] Zenodo, "Zenodo's infrastructure," 2018. [Online]. Available: http://about.zenodo.org/ infrastructure/

[30] DSpace, "Sitio oficial," 2018. [Online]. Available: https://dspace.org/

[31] OpenDOAR, "The directory of open access repositories," 2018. [Online]. Available: http:www.opendoar.org

[32] J. Lasser, "Creating an executable paper is a journey through open science," *Communications Physics*, vol. 3, no. 1, pp. 1–5, 2020, doi: 10.1038/s42005-020-00403-4

[33] N. Vasilevsky, J. Minnier, M. Haendel, R. Champieux, "Reproducible and reusable research: are journal data sharing policies meeting the mark?," *PeerJ*, vol. 5, p. e3208, 2017, doi: 10.7717/peerj.3208

[34] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, "Jupyter notebooks-a publishing format for reproducible computational workflows," in *ELPUB*, pp. 87–90, 2016, doi: 10.3233/978-1-61499-649-1-87

[35] "Declaración de Panamá sobre Ciencia Abierta," Dec. 2018. [Online]. Available: https: //web.karisma.org.co/wp-content/uploads/ download-manager-files/declaracion_ panama_ciencia_abierta.pdf

[36] H. Asorey, L. Núñez, C. Sarmiento-Cano, "Exposición temprana de nativos digitales en ambientes, metodologías y técnicas de investigación en la universidad," *Revista Brasileira de Ensino de Física*, vol. 40, no. 4, 2018m doi: 10.1590/1806-9126-rbef-2018-0092

[37] R. Mayo-García, L. Nuñez, H. Asorey, M. Rodriguez-Pascual, D. Cazar Ramirez, L. A. Torres-Nino, "Data Accessibility, Reproducibility and Trustworthiness with LAGO Data Repositories," in *Proceedings of The 34th International Cosmic Ray Conference — PoS(ICRC2015)*, 2016, p. 672, doi: 10.22323/1.236.0672

[38] E. Barrios, R. Torréns, L. Torres, L. Núñez, "Implementación de un repositorio de datos científicos usando dspace," 2011. [Online]. Available: http://repository.urosario.edu.co/handle/10336/2465