Modelo de Gestión de Recursos Computacionales para Asistir la Reproducibilidad de Experimentos Científicos

Alexander Martínez Méndez

Trabajo de Grado para optar al título de Magíster en Ingeniería de Sistemas e Informática

Director

Dr. Luis A. Núñez

Co-Director

Dr. Gabriel R. Pedraza

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingeniería de Sistemas e Informática

Bucaramanga

2022

Dedicatoria

Dedicado a la ciencia abierta por su aporte a la democratización del conocimiento. Que esta tesis sirva como una pequeña contribución para seguir avanzando hacia un futuro en el que la información científica sea un bien público al alcance de todos.

Agradecimientos

Quiero expresar mi más profundo agradecimiento a todas las personas que me han apoyado durante mi camino en la maestría. En primer lugar, a mis padres y hermanas, Alvaro, Meby, July y Angie, por su amor incondicional, apoyo constante y confianza en mí.

A mis amigos y amigas, especialmente a Eliecer, les agradezco su apoyo constante. Gracias por estar siempre presentes. Gracias por cada momento compartido y por ayudarme a mantener la cordura en los tiempos difíciles

A mis directores de tesis, Luis y Gabriel, les agradezco por su guía, conocimiento y dedicación en cada paso de este proceso. Su infinita paciencia y enseñanzas me han ayudado a convertirme en un mejor investigador. También quiero agradecer a LA-CoNGA physics y al grupo Halley por brindarme la oportunidad de trabajar con personas maravillosas en un entorno colaborativo y estimulante.

Por último, quiero agradecer a la universidad por brindarme las herramientas y el conocimiento necesarios para alcanzar este logro académico. Gracias por crear un ambiente de aprendizaje enriquecedor, lleno de oportunidades para crecer y desarrollarme como profesional.

Tabla de Contenido

1. Introducción **16**

2. Pregunta de Investigación **18**

3. Objetivos **20**

3.1. Objetivo general 20

3.2. Objetivos específicos 20

4. Metodologías para la Reproducibilidad **21**

4.1. Documentación 21

4.2. Gestión de Datos 22

4.3. Gestión de Software 24

4.4. Gestión de Entornos de cómputo 24

4.5. Gestión de Comunicaciones 25

4.6. Licencias y Derechos de Uso 26

5. Metodología **28**

5.1. Caracterización 28

5.1.1. Caracterización de elementos metodológicos para la reproducibilidad 28

5.1.2. Definición de Tecnologías y Protocolos del Modelo 29

5.2. Diseño del Modelo 30

5.3. Evaluación del modelo 30

6. Modelo de Gestión de Recursos para la Reproducibilidad de Experimentos 30

6.1. Planteamiento del experimento 32

6.2. Análisis 33

6.2.1. Gestión de datos 33

6.2.2. Gestión de software 35

6.2.3. Gestión de Métodos 37

6.3. Información volátil 38

6.4. Publicación 38

7. Validación del Modelo 39

7.1. Casos de uso 39

7.1.1. Casos de uso consolidados 40

7.1.2. Casos de uso en consolidación 43

7.2. Estadísticas de uso de la plataforma MiLab 45

7.2.1. Servicio *chatLab* 45

7.2.2. Servicio *G-Lab* 45

7.2.3. Servicio *dataLab* 46

8. Conclusiones 47

Referencias Bibliográficas **52**

Apéndices **60**

Lista de Figuras

Figura 1. Definición de reproducibilidad por *The Turing Way* basada en el uso de los análisis y datos del experiento. Adaptado de Community et al. (2019). 19

Figura 2. Abstracción del flujo de trabajo en la producción científica. De los laboratorios a los reportes de resultados. 20

Figura 3. Aspectos a tener en cuenta para ser exitosos en la gestión de datos de investigación. Tomado de de Waard et al. (2015). 22

Figura 4. Modelo metodológico para mejorar la reproducibilidad de experimentos científicos. 31

Figura 5. Captura de pantalla del tablero tipo *canvan* usado por el *grupo Halley* para la gestión de actividades en el marco del programa *LA-CoNGA physics*. 42

Figura 6. Estadísticas generales de uso del servicio *chatLab*. 45

Figura 7. Estadísticas de envío de mensajes en el servicio *chatLab* para el periodo 2022-07-17 a 2022-08-16. 46

Figura 8. Comportamiento del total de usuarios del servicio *G-Lab* en el periodo julio-2021 a julio-2022. 46

Figura 9. Comportamiento del total de proyectos del servicio *G-Lab* en el periodo septiembre-2021 a agosto-2022. 47

Figura 10. Comportamiento del total de pipelines (acciones CI/CD) del servicio *G-Lab* en el periodo septiembre-2021 a agosto-2022. 47

Figura 11. Comportamiento del total de incidencias y Merge requests del servicio *G-Lab* en el periodo septiembre-2021 a agosto-2022. 48

Figura 12. Dataset publicados en el servicio *dataLab*. 48

Figura 13. Ruta para obtener publicaciones con alto grado de reproducibilidad por *The Turing Way*. Adaptado de Community and Scriberia (2020). 60

Figura 14. Representación tipo árbol del historial de cambios en un proyecto gestionado con control de versiones. Adaptado de *storyset on Freepik*. 62

Figura 15. Arquitectura de contenerización (izquierda) vs. virtualización (derecha). 63

Figura 16. Arquitectura del servicio `https://mybinder.org/`. Tomado de *The Turing Way Scriberia* Community and Scriberia (2020). 64

Figura 17. Arquitectura del ecosistema RSpace para el soporte a la reproducibilidad y los datos FAIR. Tomado de `https://www.researchspace.com/` 66

Figura 18. Arquitectura de los Notebooks Jupyter. Adaptado de `https://securitydatasets.com/consume/jupyter-notebooks.html` 66

Figura 19. Servicios de la plataforma para análisis interactivos en la nube *SWAN*. Adaptado de Piparo et al. (2018). 67

Figura 20. Plataforma MiLab para el apoyo a la investigación y el control de proyectos. 69

Figura 21. Flujo de trabajo en entornos de investigación mediante los servicios *MiLab*. 70

Figura 22. Arquitectura de autenticación federada para los servicios *MiLab*. 73

Figura 23. Diagrama de organización de Dataverses en el servicio *dataLab*. Adaptado de `https://guides.dataverse.org/en/4.5/user/dataverse-management.html` 75

Figura 24. Diagrama de organización de Datasets en el servicio *dataLab*. Adaptado de `https://guides.dataverse.org/en/latest/user/dataset-management.html` 75

Figura 25. Taxonomía de la Ciencia Abierta. Adaptado de `https://doi.org/10.1145/2809563.2809571` 78

Lista de Tablas

Tabla 1. Resumen de casos de uso para la validación del modelo 44

Tabla 2. Especificaciones de cómputo de la plataforma *MiLab* 77

Lista de Apéndices

		pág.
Apéndice A.	Herramientas para la reproducibilidad	60
Apéndice B.	Plataforma MiLab	68
Apéndice C.	Ciencia Abierta	77

Resumen

Título: Modelo de Gestión de Recursos Computacionales para Asistir la Reproducibilidad de Experimentos Científicos [*]

Autor: Alexander Martínez Méndez [**]

Palabras Clave: Reproducibilidad, Ciencia Abierta, Acceso Abierto, Metodología, Plataforma, MiLab, Reproducible, Replicable, Repetible

Descripción: La reproducibilidad, entendida como la capacidad de replicar o repetir los procesos de investigación, es la característica que permite validar el conocimiento producido en la actividad científica. Sin embargo, la reproducibilidad de la ciencia está en crisis. Diversos trabajos, encuestas y autores lo han evidenciado, especialmente desde el movimiento de Ciencia Abierta. El contexto tecnológico actual y las nuevas formas de hacer ciencia, el crecimiento exponencial en la producción de datos y el mismo sistema de medida de productividad de la comunidad académica, son solo algunas de las causas de esta crisis.

Han surgido diversas y numerosas propuestas para resolver cada uno de los desafíos generados por la reproducibilidad de la ciencia, con el respaldo de la Ciencia Abierta. A pesar de esto, solucionar la crisis en la reproducibilidad requiere aún de grandes esfuerzos. El enfoque en el desarrollo de herramientas tecnológicas, más allá de su utilidad y calidad, no parece tener grandes efectos. Se debe abordar desde un enfoque integral, donde lo tecnológico vaya de la mano de las metodologías que soportan su uso. Las soluciones propuestas deben apoyar los procesos de investigación desde el planteamiento de la pregunta de investigación, hasta la publicación de los resultados. Es indispensable también la formación del personal de investigación en prácticas que fomenten la reproducibilidad.

---

[*] Trabajo de Maestría

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Luis A. Nuñez. Co-Director: Gabriel R. Pedraza

En este trabajo presentamos un modelo metodológico para asistir y mejorar la reproducibilidad de experimentos científicos en pequeños y/o medianos grupos de investigación. Su desarrollo giró en torno a tres acciones indispensables para la reproducibilidad: preservar, compartir y colaborar. El modelo ofrece rutas de acción para la gestión de datos, códigos computacionales, entornos software y la información volátil en experimentos (sus comunicaciones). La evaluación se desarrollo mediante la implementación de casos de uso, sobre una plataforma tecnológica abierta creada en el marco de este proyecto.

El modelo y la plataforma tecnológica desarrollada permitieron crear entornos de investigación con la reproducibilidad como eje principal. Los procesos de preservación y documentación pasaron de ser una lista de entregables al final de una publicación científica, a ser una labor diaria en la práctica investigativa.

Abstract

Title: Computational Resource Management Model to Support the Reproducibility of Scientific Experiments [*]

Author: Alexander Martínez Méndez [**]

Keywords: Reproducibility, Open Science, Open Access, Open Access, Methodology, Platform, MiLab, Reproducible, Replicable, Repeatable

Description: Reproducibility, understood as the ability to replicate or repeat research processes, is the characteristic that allows the validation of the knowledge produced by scientific activity. However, reproducibility of science is in crisis. Various works, surveys and authors have demonstrated this, especially since the Open Science movement. The current technological context and new ways of doing science, the exponential growth of data production, and the same productivity measurement system of the academic community are just some of the causes of this crisis.

Numerous proposals have emerged to address each of the challenges posed by the reproducibility of science, supported by Open Science. However, solving the reproducibility crisis still requires considerable efforts. Focusing on the development of technological tools, beyond their usefulness and quality, does not seem to have a significant impact. It needs to be addressed from a comprehensive approach, where technology goes hand in hand with the methodologies that support its use. The solutions proposed should support research processes from the formulation of research questions to the publication of results. It is also essential to train research staff in practices that promote reproducibility.

In this work, we present a methodological model to support and improve the reproducibility of scientific experiments in small and/or medium-sized research groups. Its development revolved around three essential actions for reproducibility: preserve, share and collaborate. The model offers action paths for data management, computational codes,

---

[*] Master's thesis

[**] Faculty of Physical-Mechanical Engineering. School of Systems and Computer Engineering. Director: Luis A. Núñez. Co-Director: Gabriel R. Pedraza.

software environments and volatile information in experiments (their communication). The evaluation was carried out through the implementation of use cases on an open technological platform created within this project.

The model and the technological platform developed allowed the creation of research environments with reproducibility as the main focus. The processes of preservation and documentation were transformed from a list of deliverables at the end of a scientific publication to a daily task in research practice.

1. Introducción

La producción de conocimiento científico se da bajo ciertas premisas que buscan dar confiabilidad al trabajo y los resultados del proceso de investigación. La reproducibilidad, entendida como la capacidad de replicar o repetir los procesos de investigación científica, es la característica que permite validar los resultados expuestosGoodman et al. (2016). Es también esta característica la que permite identificar cuestionables prácticas en estos procesos. Sin embargo, el contexto tecnológico actual y las nuevas formas de hacer ciencia, el crecimiento exponencial en la producción de conjuntos de datos y el mismo sistema de medida de productividad de la comunidad académica, han generado una crisis en la reproducibilidad de la actividad científica.

La afirmación de una crisis en la reproducibilidad la soporta en gran medida dos encuestas. Una realizada por la revista *Nature* a más de 1500 investigadores e investigadoras, donde el 90 % de las personas encuestadas reconocen la crisis Baker (2016). Y otra encuesta entre personal de investigación de los Países Bajos dónde se evidencian malas conductas en la investigación y se indaga sobre los factores que determinan la incidencia en estas conductas Vrieze (2021). Se suman a estas encuestas las voces del personal académico en artículos editoriales y se evidencia esta crisis en el número de retractaciones reportadas en el portal *Retraction Watch*.

Muchas propuestas han surgido para dar solución a cada uno de los desafíos generados en la reproducibilidad de la ciencia, gran parte desde el creciente movimiento de Ciencia Abierta. Un ejemplo del impulso desde este movimiento es la guía de la Oficina Estadounidense de Política Científica y Tecnológica de la Casa Blanca (OSTP por sus siglas en inglés) para que la investiga-

ción financiada por el gobierno federal esté disponible sin periodos de embargoNelson (2022). A pesar de esto, el escenario actual nos muestra que hay un largo camino por recorrer para la solución de la crisis de la reproducibilidad. Se debe explorar soluciones con un enfoque integral, ir más allá de ofrecer herramientas para la gestión de determinado elemento. Las soluciones propuestas deben apoyar los procesos de investigación desde el planteamiento de la pregunta de investigación, hasta la publicación de los resultados con todos los elementos que los soportan. Además, la formación del personal de investigación en prácticas que fomenten la reproducibilidad es indispensable, en este sentido organizaciones como *The Turing Way*Community et al. (2019), el programa *Software Carpentry* Wilson (2016) o la red *UK Reproducibility Network (UKRN)* Committee (2021) han realizado un extraordinario trabajo.

En este trabajo presentamos un modelo metodológico para asistir y mejorar la reproducibilidad de experimentos científicos en pequeños y/o medianos grupos de investigación. Aquellos de hasta 50 personas aproximadamente y donde se hace más dificil destinar recursos para este tipo de labores. Entendiendo la reproducibilidad como un proceso, planteamos un conjunto de prácticas desde la concepción de los experimentos, hasta la presentación de los resultados para lograr un alto grado de reproducibilidad. El modelo ofrece rutas de acción para la gestión de datos, códigos computacionales, entornos software y la información volátil del experimento (sus comunicaciones en servicios de mensajería). El modelo se validó mediante el desarrollo de una plataforma

tecnológica para soportar los procesos sugeridos en diversos casos de uso.

2. Pregunta de Investigación

¿Es posible mejorar la crisis de reproducibilidad de la ciencia mediante el uso de herramientas y metodologías computacionales en los procesos de investigación?

La reproducibilidad es el elemento fundamental en nuestra pregunta de investigación, validar el conocimiento generado es posible si se puede reproducir, replicar o repetir cada uno de los procesos sobre los cuales se desarrolló el experimento. Sin embargo, la reproducibilidad de la ciencia está en crisis, diversos trabajos se han realizado sobre el tema, muchos autores lo han expresado abiertamente Munafò et al. (2017); Buck (2015); Perkel (2020); Barba (2016); Peng (2015); Stodden et al. (2016); DMA (2019); Andrew (2018); Yang et al. (2020); Camerer et al. (2018); Mondelli et al. (2019); Markowetz (2015); Dirnagl (2020); Allison et al. (2016); Peng (2011); Fanelli (2018). Se han reconocido también algunas de las barreras para lograr la reproducibilidad Whitaker (2018).

Cabe mencionar que en el transcurso de este documento se supone la reproducibilidad de manera global, es decir, reproducibilidad, replicabilidad y repetibilidad. Otras definiciones sobre la reproducibilidad se basan en los datos, métodos y resultados como la propuesta por Steven N. Goodman, et al, en *"What does research reproducibility mean?"* Goodman et al. (2016), la propuesta por *The Turing Way* [Ver figura 1] o el espectro de la reproducibilidad propuesto por Peng R. et al, en *Reproducible Research in Computational Science* Peng (2011).

La solución a la crisis implica reconocer la reproducibilidad como un proceso desde la concepción del experimento y no como un ítem más en una lista de entregables. Para esto, lo

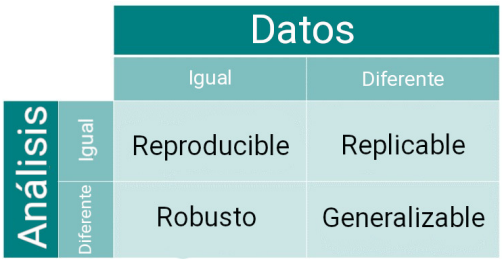

Figura 1. Definición de reproducibilidad por *The Turing Way* basada en el uso de los análisis y datos del experiento. Adaptado de Community et al. (2019).

tecnológico debe apoyarse en la implementación de metodologías que permitan la adopción y correcto uso de las herramientas o aplicaciones en cada uno de los procesos de investigación.

Existen herramientas de gran calidad y con las funcionalidades necesarias para gestionar los elementos propios de la reproducibilidad [Ver apéndice 1]Leisch et al. (2011); Pain (2018); Xie and Allaire (2012); Brammer et al. (2011); Ivie and Thain (2016); Goodman et al. (2017), sin embargo, su adopción no se ha dado de la mejor manera, esencialmente por la ausencia de metodologías que permitan su integración a los procesos. Este trabajo de investigación plantea un modelo Gestión de Recursos Computacionales para Asistir la Reproducibilidad de Experimentos Científicos con un enfoque metodológico.

El modelo busca acoplarse a los flujos de trabajo en los procesos de investigación desde la fuente de los datos hasta la publicación de los resultados [Ver figura 2]. Específicamente el modelo plantea consideraciones desde la captura, adquisición, creación o recolección de conjuntos de datos hasta su preservación; la gestión de códigos computacionales ; los análisis en entornos de cómputo; y la documentación sobre los procesos, protocolos y el proyecto en general.

Figura 2. Abstracción del flujo de trabajo en la producción científica. De los laboratorios a los reportes de resultados.

3. Objetivos

3.1. Objetivo general

Proponer un modelo de gestión de recursos computacionales apoyado en una plataforma de software, para asistir la reproducibilidad de experimentos científicos de cálculo y/o simulación, enmarcados en el proyecto La-CoNGA physics.

3.2. Objetivos específicos

Caracterizar los recursos computacionales, funcionalidades y protocolos de interoperabilidad en contextos de grupos de investigación de pequeña y mediana escala para su inclusión en el modelo.

Diseñar modelo de gestión de recursos computacionales basado en una plataforma de software.

Validar el modelo propuesto mediante la implementación de un prototipo enmarcado en el proyecto La-CoNGA physics.

4. Metodologías para la Reproducibilidad

De manera general mencionamos la Ciencia Abierta [Ver apéndice 3] y especialmente las tendencias de acceso abierto que se vienen dando desde este movimiento. La Ciencia Abierta permite desarrollar experimentos más transparentes, éticos y reproducibles desde la apertura de todos los elementos del proceso de investigación. *The Turing Way*Community et al. (2019) distingue tres niveles para la investigación abierta, el primero de ellos y el más importante, el accesos sin restricciones a los elementos de la investigación. En el segundo nivel se tiene la reusabilidad, acá se destacan los protocolos y estándares que permitan el uso e interpretación de los elementos publicados. En el tercer y último nivel está la transparencia, los metadatos y la documentación son fundamentales para conocer cómo, bajo qué condiciones y qué o quiénes intervinieron en la creación de cada elemento del experimento.

A continuación se presenta un recorrido a través de las prácticas metodológicas recomendadas para permitir y mejorar la reproducibilidad de proyectos de investigación.

4.1. Documentación

La documentación es esencial para la reproducibilidad, es el punto inicial para reproducir cualquier experimento, en cualquier escenario. Se deben agregar datos específicos a cada elemento para facilitar su entendimiento, incluso de sus propios autores. Destacamos tres elementos de documentación relevantes para la reproducibilidad de los experimentos.

- Los metadatos, entendidos como datos sobre datos, describen el ciclo de vida de los elementos en una investigación. Es indispensable el uso de estándares que permitan una correcta

comunicación entre sistemas Gregg et al. (2019), además del uso de vocabularios y diccionarios que garanticen el entendimiento de cada uno de los metadatos asociados.

- Las buenas prácticas de programación para la documentación de códigos computacionales.
- Finalmente, la documentación sobre los protocolos, métodos, instrumentos y todos los elementos usados en el proceso de investigación.

4.2. Gestión de Datos

Suponiendo los datos como el insumo principal para los experimentos, la gestión de los mismos se vuelve primordial y se debe hacer de la mejor manera. de Waard, et al, presentan una hoja de ruta en la que se exponen las características de mayor importancia a la hora de desarrollar políticas de gestión de datos exitosas de Waard et al. (2015).

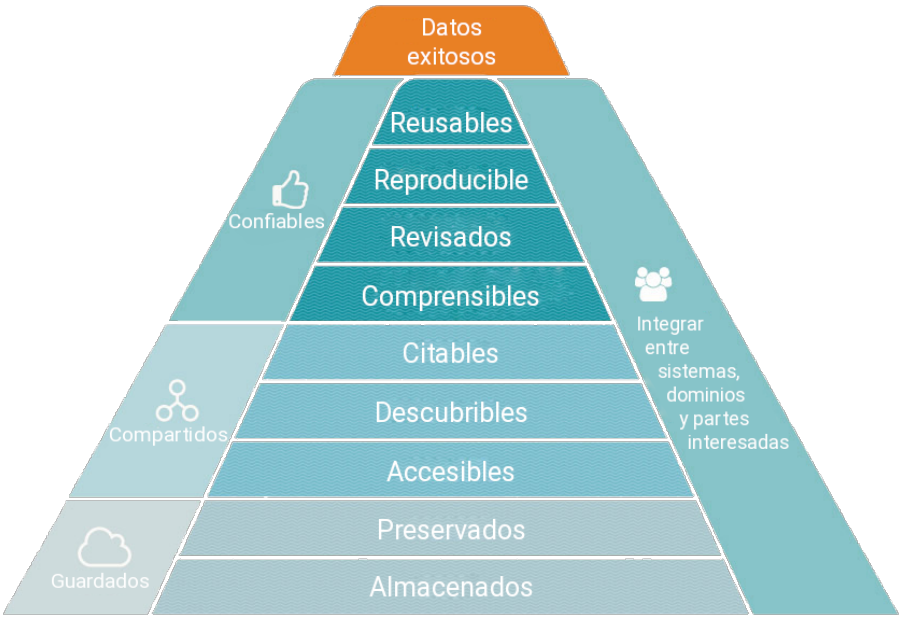

Figura 3. Aspectos a tener en cuenta para ser exitosos en la gestión de datos de investigación. Tomado de de Waard et al. (2015).

En la hoja de ruta propuesta por de Waard, et alde Waard et al. (2015), a grandes rasgos se destacan los siguientes aspectos [Ver figura 3]. El almacenamiento, la preservación y el acceso de los conjuntos de datos, especialmente en entornos de ciencia abierta. La confiabilidad, tanto de los datos en sí mismos, como de los métodos e instrumentos usados para su recopilación. Finalmente, todas las características deben integrarse para completar el ciclo de gestión de datos.

Diversos autores han dado, desde su experiencia, recomendaciones para la gestión de datos, recomendaciones que bien podría decirse giran en torno a los aspectos expuestos en la figura 3. Entre estos encontramos a Haak W.Haak (2019) con sus 4 principios para desbloquear el potencial de los datos de investigación. Delevante R.Delevante (2019) en representación del comité asesor para la gestión de datos en la organización Mendeley, con su exposición de las 5 tendencias para la gestión de datos. Willems L.Willems (2019) y las 6 lecciones aprendidas desde el piloto para apoyar la gestión de datos de investigación en 4 instituciones educativas. El llamado para generar una infraestructura que permitiera explotar todo el potencial de los datos para la ciencia y la sociedad expuesto por van der Graaf M.van der Graaf and Waaijers (2011) en A Surfboard For Riding The Wave. Ailamaki A.Ailamaki et al. (2010) en su exploración de los requerimientos principales en sistemas de gestión de datos de propósito general. El CWTS (Centre for Science and Technology Studies) for Science and Technology Studies (2017) con el reporte sobre la perspectiva de datos abiertos desde la mirada de personas en el mundo de la investigación. Finalmente, Wilkinson M. et al, Wilkinson et al. (2016) con los comentarios sobre los principios FAIR (Fair, Accesible, Interoperable, Reusable) para la gestión de datos científicos.

En términos metodológicos para la gestión de datos los *Planes de Gestión de Datos*, DMP

(por sus siglas en inglés), son el punto de partida para una estrategia de gestión de datos. Un DMP es un documento en el cual se definen todas las características y acciones de gestión de los datos a usarse en el experimento Dcc (2013). Los DMP evolucionarán con el desarrollo del experimento y además de la relevancia para la reproducibilidad del experimento, de manera recurrente hacen parte de los entregables a los fondos de financiación.

El proyecto LEARN desde plantillas y casos de estudio nos presenta una completa guía para la gestión de datos de investigación noa (2017). Decisiones metodológicas, administrativas y tecnológicas son abordadas mediante casos de estudio y la política de gestión de datos de investigación en University of Vienna.

4.3. Gestión de Software

Las piezas de software o códigos computacionales permiten extraer información de los conjuntos de datos y evaluar o construir posibles escenarios en el marco de una investigación. El uso de sistemas de control de versiones es la herramienta metodológica más adecuada para la gestión de este elemento, con sistemas de este tipo se logra trazabilidad y preservación, además de facilitar el trabajo colaborativo. Sumado a esto, se debe tener en cuenta también el uso de buenas prácticas de programación Wilson et al. (2017); Chalstrey (2021); Van Atteveldt et al. (2019) o guías de estilo Piater (2005) que permitan la adecuada interpretación de la pieza de software.

4.4. Gestión de Entornos de cómputo

Los entornos de cómputo en investigación involucran tanto el hardware como el software sobre el que se ejecutan las piezas de software o códigos computacionales. El entorno puede influir en la reproducibilidad de un experimento de diversas maneras. Entre estas, variaciones en

resultados al usar entornos diferentes o la imposibilidad de construir un entorno con la capacidad de realizar los análisisPerkel (2020). Idealmente para la reproducibilidad la mejor opción siempre será usar el mismo entorno de cómputo, sin embargo esto es algo difícil de lograr. Metodológicamente es más eficiente encapsular el entorno de software mediante tecnologías de contenerización o virtualizaciónFelter et al. (2015); Poldrack et al. (2019); Trisovic (2018), documentando las características del software y especialmente el hardware original. A mayor cantidad de información sobre el entorno de cálculo, mayor probabilidad de lograr una correcta ejecución del software. La arquitectura del hardware usado, las dependencias del software (librerías, plugins, dispositivos, etc) o simplemente la versión usada, son solo algunos de los elementos a documentar y preservar para permitir la reproducibilidad del experimento.

4.5. Gestión de Comunicaciones

Las discusiones en torno al desarrollo de un proceso de investigación es uno de los factores menos abordados cuando se habla de la crisis de la reproducibilidad y sus posibles causas. Sin embargo, es en estas discusiones donde esta la historia de un experimento, allí están los detalles de cada acción tomada en el transcurso de una investigación y son un factor de gran importancia para la reproducibilidad. La centralización y preservación de las discusiones del proceso de investigación son los elementos más relevantes para la reproducibilidad. El uso de servicios de mensajería instantánea (Whatsapp, Telegram, Messenger, etc.), Redes sociales o el correo electrónico, aunque altamente usados, actúan en contra de estos dos elementos, la información se dispersa y generalmente se pierde cuando las personas abandonan los grupos o centros de investigación.

Centralizar la discusiones permite una interacción más fluida y organizada, además de faci-

litar la búsqueda de información y su preservación. La preservación por su lado, permite el acceso a la información en cualquier momento. Se destaca en este elemento el uso de herramientas de comunicación que permitan la organización de las comunicaciones por temas de interés y que permitan comunicaciones tanto síncronas como asíncronas.

4.6. Licencias y Derechos de Uso

Finalmente en este recorrido por las prácticas y elementos para la reproducibilidad encontramos los derechos de uso, específicamente las licencias. En estas están descritas las condiciones legales de uso del elemento al cual estamos accediendo o compartiendo. Por defecto cualquier elemento publicado o compartido tiene restricciones de uso, copia, modificación y redistribución, solo podrá realizarse alguna de estas acciones con autorización explícita del autor(a). Mediante la asignación de licencias libres se puede permitir de manera clara el acceso, uso, copia, modificación, redistribución o cualquier subconjunto de estas. El uso de licencias libres Ballhausen (2019) o de mejor manera, el dominio público, son el mejor escenario para la reproducibilidad. Existe un gran número de licencias abiertas usadas de manera extendida, en `https://opensource.org/licenses/category` se puede consultar un listado generado por la Open Source Initiative (OSI).

El panorama de la reproducibilidad es prometedor desde el reconocimiento de la crisis actualBaker (2016), sin embargo, los desafíos a resolver requieren aún de grandes esfuerzos. Aunque los problemas en la reproducibilidad de experimentos no es una noticia reciente, en la pasada década estos han sido el tema de estudio de una gran cantidad de publicaciones, trabajos de investigación y especialmente editoriales en revistas de alto impacto. Desde iniciativas para proponer

soluciones colectivamente, hasta grandes encuestas para identificar los puntos claves en la crisis de la reproducibilidad han sido publicados. Solucionar los problemas de reproducibilidad es más que la implementación de herramientas, la integración de éstas a los procesos de investigación es el elemento diferenciador. La formación del personal de investigación en prácticas que fomenten la reproducibilidad es indispensable y un gran trabajo han realizado en este sentido organizaciones como *The Turing Way*Community et al. (2019), el programa *Software Carpentry* Wilson (2016) o la red *UK Reproducibility Network (UKRN)* Committee (2021).

Los avances para solventar la crisis de reproducibilidad son más visibles desde lo práctico [Ver apéndice 1 para una descripción más detallada del estado del arte de las herramientas software para la reproducibilidad], esto desde el desarrollo de herramientas software para apoyar la reproducibilidad. Sin embargo, y siendo este uno de los puntos clave para este trabajo de investigación, la implementación de herramientas tecnológicas sin los elementos metodológicos adecuados, limitan sus capacidades para resolver los problemas de reproducibilidad.

Para cerrar este capítulo, mencionamos la importancia de reconocer la investigación o el conjunto de prácticas que permiten hacer reproducible la cienciaKoers (2016); Ioannidis (2014); Nosek et al. (2015); Begley and Ioannidis (2015). Para esto se requiere el compromiso de todas las partes en la cadena de producción de conocimiento y de muchos cambios en el sistema académico. Es esencial cambiar los sistemas de medición y evaluación para fomentar las prácticas asociadas a

la reproducibilidad y/o la Ciencia Abierta.

5. Metodología

En este capítulo se describe la metodología utilizada para el desarrollo de este proyecto de investigación. El proyecto se realizó en 3 etapas, caracterización, diseño y evaluación. En la etapa de caracterización se hizo una búsqueda y análisis de los elementos metodológicos más relevantes para la reproducibilidad de un experimento. Se indagó en la literatura y en centros de investigación, sobre las prácticas para la mejora de la reproducibilidad de sus experimentos. En la etapa de diseño se recogieron los elementos de la etapa anterior y se usaron para el desarrollo del modelo de gestión de recursos computacionales. En la etapa final se evaluó el modelo propuesto mediante la implementación de casos de uso sobre la plataforma MiLab.

5.1. Caracterización

Esta etapa consistió en una revisión de literatura para caracterizar los procesos con influencia directa en la reproducibilidad de los experimentos. Se identificaron también las herramientas tecnológicas asociadas a estos procesos, esto para su uso en la implementación y evaluación del modelo. Destacamos el proyecto *The Turing Way*Community et al. (2019), un aglomerado de recomendaciones (de los más completos en el área) para realizar ciencia de datos *reproducible, ética y colaborativa*. Aunque su enfoque es la ciencia de datos, sus contenidos pueden generalizarse para el fomento de la reproducibilidad de cualquier experimento.

5.1.1. Caracterización de elementos metodológicos para la reproducibilidad.

En la primera etapa de este trabajo identificamos y generalizamos los procesos de investigación. Indagamos en la literatura, especialmente en las enseñanzas obtenidas por grandes grupos de inves-

tigación Community et al. (2019); Manghi et al. (2010); Poldrack et al. (2019); Piparo et al. (2018); Willems (2019); Kikkenborg (2019); Feger et al. (2019); Florez Vargas (2016); Chen et al. (2019), las buenas prácticas y metodologías asociadas a cada proceso para favorecer su reproducibilidad. Revisamos principalmente los siguientes factores.

- Área de investigación
- Flujos y protocolos de trabajo.
- Recursos digitales involucrados.
- Acceso abierto.
- Vigencia de las metodologías propuestas.
- Factibilidad de implementación.

5.1.2. Definición de Tecnologías y Protocolos del Modelo. A partir de los procesos y metodologías identificadas en la etapa anterior, realizamos un análisis de mercado sobre las tecnologías y herramientas asociadas a estas. Se hizo necesario este análisis para la implementación de los casos de uso en la evaluación del modelo. Los parámetros revisados fueron los siguientes.

- Capacidad de uso.
- Capacidad de integración.
- Capacidad de adaptación.

- Estar bajo software libre o abierto.
- Escalabilidad.

5.2. Diseño del Modelo

En esta etapa diseñamos el *modelo de gestión de recursos computacionales para asistir la reproducibilidad de experimentos científicos*. El modelo se desarrolló con la premisa de que este debe acoplarse a los procesos de investigación de inicio a fin para impactar positivamente en la reproducibilidad del experimento. El diseño del modelo se realizó de manera iterativa, abordando en cada iteración las etapas de planteamiento, análisis y publicación de un experimento de investigación. El modelo desarrollado se encuentra descrito en el capítulo 6 de este documento.

5.3. Evaluación del modelo

El modelo desarrollado se evaluó mediante la implementación de 9 casos de uso sobre la plataforma MiLab (plataforma descrita en el apéndice 2 y desarrollada en el marco de este trabajo). La evaluación se encuentra en el capítulo 7 de este documento.

6. Modelo de Gestión de Recursos para la Reproducibilidad de Experimentos

En este capítulo se describe el modelo desarrollado y los aspectos metodológicos tenidos en cuenta para su diseño. La reproducibilidad es entendida como un proceso y para esto, planteamos un conjunto de prácticas desde la concepción del experimento, hasta la presentación de los resultados para un experimento reproducible. El punto de quiebre para tener experimentos reproducibles está en los aspectos metodológicos y el uso de herramientas software es solo un elemento técnico en los procesos de investigación.

En la figura [4] se presenta de manera gráfica el modelo metodológico desarrollado para mejorar la reproducibilidad de los experimentos. El modelo es una generalización del flujo de trabajo para experimentos con alto grado de reproducibilidad. Se menciona "grado de reproducibilidad.^entendiendo que la reproducibilidad no es una característica binaria, en algunos casos y generalmente por cuestiones técnicas, solo fragmentos o análisis a menor escala pueden ser reproducidos. El modelo propuesto busca mejorar la reproducibilidad con la curación, preservación y documentación de la mayor cantidad de elementos presentes en el desarrollo del experimento.

Figura 4. Modelo metodológico para mejorar la reproducibilidad de experimentos científicos.

El modelo presentado recoge los aspectos metodológicos desde su concepción, hasta la presentación de los resultados. Se parte de una primera etapa de *planteamiento* o planificación del experimento científico, donde se definen los primeros elementos relevantes para la reproducibilidad del experimento. A continuación, en la etapa de *análisis* se aborda el desarrollo del experimento desde la captura de datos, hasta su análisis e interpretación. En la tercera y última etapa se plan-

tean las consideraciones para la etapa de transferencia y publicación del conocimiento generado. Se agrega también la información volátil (Las comunicaciones sobre el desarrollo del experimento) como un elemento transversal a las anteriores etapas y el cual posee la información sobre las actividades del experimento, las decisiones tomadas y el por qué de estas. A continuación profundizaremos en cada uno de los aspectos del modelo.

6.1. Planteamiento del experimento

En la primera etapa del experimento se debe generar el *plan de gestión de datos*, allí se deben definir todas las características y acciones de gestión de los datos a usarse en el experimento. Es importante tener en cuenta que un *plan de gestión de datos* es un documento que evolucionará con el desarrollo del experimento. Sin embargo, las primeras versiones influyen de manera significativa en la reproducibilidad del experimento, especialmente por las decisiones tomadas en torno al acceso y preservación de los datos.

El *plan de gestión de datos* creado puede apoyarse en una gran cantidad de modelos/plantillas disponibles, incluso herramientas tipo formulario disponibles en la web, teniendo en cuenta que este debe responder a las características específicas del experimento. De manera general y usando el modelo propuesto por Dcc (2013) se debe incluir la siguiente información en el *plan de gestión de datos*.

- Datos administrativos. Nombre y descripción del experimento, fondos de financiación, el centro de investigación o personas involucradas, entre otros.
- Captura de datos. Información sobre qué datos serán capturados y la forma de captura.

- Documentación y metadatos: Definiciones sobre los metadatos asociados a los conjuntos de datos y cualquier otro tipo de documentación sobre estos.
- Aspectos éticos y legales. Información y definiciones sobre los derechos de uso (Licencia), protección de datos personales y consentimiento informado sobre su uso.
- Almacenamiento. Información de la forma de almacenamiento y acceso de los datos, además de las consideraciones de seguridad para el restablecimiento de los mismos.
- Preservación. Definiciones sobre los criterios de preservación, conjuntos de datos a preservar y el tiempo de preservación.
- Compartir. Información sobre la forma de compartir los datos y las restricciones de acceso.
- Responsabilidades y recursos. Información sobre las personas a cargo de la gestión de los conjuntos de datos y los recursos necesarios para su gestión, desde la captura hasta la preservación.

6.2. Análisis

La etapa de análisis en el modelo propuesto concentra las actividades de gestión de datos, de software y de los métodos usados. A continuación se describen las consideraciones metodológicas para cada una de estas.

6.2.1. Gestión de datos. El *plan de gestión de datos* creado en la etapa anterior, es el punto de partida para todas las actividades en torno a la gestión de los conjuntos de datos. Los

aspectos a revisar en este documento para facilitar la reproducibilidad del experimento serán los siguientes.

- Formato de los datos. Se debe usar formatos abiertos que permitan su uso y lectura sin restricciones o requisitos de software. Usar por ejemplo *CSV* en lugar de los formatos *Microsoft Excel*.

- Licencia. La licencia usada o generada para definir los derechos de los conjuntos de datos debe permitir, su uso y modificación son restricciones. Se recomienda usar el dominio público para datos de investigación.

- Metadatos. Se debe revisar si los metadatos permiten mínimo interpretar los conjuntos de datos y el contexto de su captura.

Revisados estos elementos, encontramos en la guía "10 aspects of highly effective research data"propuesta por de Ward A. et al. de Waard et al. (2015) y presentada en el capítulo anterior [Ver figura 3], las acciones a seguir para que la gestión de los conjuntos de datos permita la reproducibilidad del experimento. A continuación las acciones a realizarse.

- Los datos deben almacenarse y preservarse garantizando su integridad, esto sin distinción de ser datos medidos o simulados. Algunas características de los experimentos asociadas a las condiciones geográficas, de conectividad, de los instrumentos, del software, entre otras, pueden influir en este proceso y deberán ser evaluadas.

- Las herramientas usadas para el almacenamiento y/o preservación deben permitir que los datos sean buscados, accedidos y citados. Se debe tener en cuenta los niveles de privacidad o tiempos de embargo de acuerdo a las características del experimento.
- Los datos deben ser confiables garantizando su compresión, revisión, reproducción y reuso.
- Finalmente, se deben usar protocolos de descubrimiento o transferencia que permitan su interoperabilidad.

Estas acciones pueden resumirse en el cumplimiento de los principios FAIR Wilkinson et al. (2016); Mondelli et al. (2019), con estos se busca garantizar que los conjuntos de datos sean encontrables, accesibles, interoperables y reusables. En términos técnicos, las herramientas software más adecuadas para la gestión de datos son los repositorio de datos de investigaciónNature (2021).

6.2.2. Gestión de software. El modelo propuesto divide el software usado en experimentos científicos en dos grupos, *software de análisis* para la captura o tratamiento de los datos y el *entorno software de cómputo* donde se ejecuta el software de análisis. A continuación se trata con detalle los elementos tenidos en cuenta para la gestión del software.

6.2.2.1. Software de análisis. La gestión del software de análisis tiene requerimientos similares a la gestión de datos. De manera general, se busca que el software usado en los análisis o procesos de captura de datos pueda ser encontrado, interpretado y re-usado. A continuación los elementos o acciones planteadas por para la gestión del software de análisis.

- Usar buenas prácticas de programaciónWilson et al. (2017); Chalstrey (2021); Van Atteveldt

et al. (2019). Desde la definición de variables hasta el diseño de módulos y funciones, deben seguir y mantener los estándares de programación propios del lenguaje.

- Documentación. Se debe describir de manera textual, clara y con detalle las acciones realizadas por el software en cada una de las etapas de procesamiento. Se debe documentar también la forma de uso y los requisitos del software.

- Trazabilidad. Se debe usar sistemas de control de versiones para generar trazabilidad de los cambios realizados a las piezas de software y registros de las personas involucradas.

- Derechos de uso. Se debe asignar una licencia que permita el uso y modificación de las piezas de software. Las licencias más adecuadas son las denominadas *licencias libres*, algunos ejemplos: GPL v3, MIT License y BSD.

- Preservar las piezas de software en sistemas especializados para garantizar su accesibilidad, uso y una forma de ser citado.

- Preservar los archivos fuente y evitar la preservación exclusiva de binarios o ejecutables que actúan como cajas negras.

Es pertinente mencionar los *Notebooks Jupyter*Thomas et al. (2016) como una de las herramientas tecnológicas más didáctica y práctica para el desarrollo de software de análisis. La integración lograda entre las piezas de software y la documentación por medio de bloques, facilita la interpretación de los mecanismos de análisis.

6.2.2.2. Entornos software de cómputo. La gestión de los entornos de software debe cumplir con al menos el primero de los siguientes elementos.

- Documentar todas las características de software y de hardware, desde las capacidades de cómputo requeridas hasta las versiones de las librerías usadas. Idealmente se espera que el escenario de cómputo pueda ser reconstruido a partir de esta información.
- Usar tecnologías de contenerización y/o virtualización para encapsular el entorno de software y permitir el despliegue del entorno de software de manera rápida sobre variadas configuraciones de hardware. Es importante mencionar que se debe mantener la documentación del entorno de hardware y software original.
- Finalmente e idealmente, ofrecer todo el entorno sobre el que se desarrollaron los análisis, esto bajo modelos tipo SaaS o similares.

6.2.3. Gestión de Métodos. El desarrollo del experimento está ligado a métodos y/o protocolos que garanticen la confiabilidad del mismo, en este sentido, deben ser documentados y preservados. La documentación debe realizarse con un alto grado de detalle, incluyendo elementos como la hora en la que se tomó una muestra o los implementos utilizados para su recolección por mencionar algunos ejemplos. Su preservación deberá permitir el acceso a la información tanto en el desarrollo del experimento como en la etapa de publicación.

Las libretas de laboratorio como medio de registro de los eventos en el desarrollo del experimento, deben ser preservadas. De manera general en la gestión de estos elementos se tienen las siguientes consideraciones.

- Usar herramientas digitales para el registro de eventos de manera estructurada.
- Digitalizar las libretas físicas (cuadernos de notas) cuando no sea posible el uso herramientas digitales.
- Preservar estos elementos junto a la documentación de los métodos y/o protocolos del experimento.

6.3. Información volátil

La centralización y preservación de las comunicaciones sobre el experimento son los elementos identificados y abordados en el modelo propuesto. Se dan a continuación las indicaciones metodológicas para la gestión de este tipo de información.

En primer lugar se debe buscar la centralización de todas las comunicaciones, síncronas o asíncronas. Establecer un canal de comunicación que reemplace los chats en servicios de mensajería instantánea, los correos electrónicos e incluso las comunicaciones en redes sociales. El canal establecido debe permitir entre otros, organizar las comunicaciones por temas de interés y búsquedas enriquecidas. Se debe tener también la capacidad de preservar en el tiempo cada mensaje, esto sin importar el posible flujo de personal en el desarrollo del experimento. En terminos prácticos los ejemplos más cercanos y populares son los servicios *Slack* y *Discord*.

6.4. Publicación

La publicación de los resultados y todo el conocimiento generado desde el desarrollo del experimento, representa la última etapa de la generalización realizada en el desarrollo de este modelo. Esta etapa es el momento de dar acceso al conjunto de elementos de las etapas anteriores.

Así, para una publicación reproducible se tendrá acceso y se permitirá el uso de los siguientes elementos.

- Los datos de investigación catalogados y preservados en formatos abiertos que permitan su reusabilidad.
- El software de análisis y la documentación del entorno software de cómputo para su ejecución.
- Documentación de los métodos y protocolos usados en el desarrollo del experimento. Además de las libretas de laboratorio usadas en cada una de las actividades del experimento.
- Las comunicaciones o discusiones en medios digitales sobre el desarrollo del experimento. Teniendo en cuenta las consideraciones de privacidad necesarias.

7. Validación del Modelo

En este capítulo se describe la validación del modelo mediante la incorporación de nueve casos de uso a la plataforma *MiLab*. Esta plataforma se describe a detalle en el apéndice 2 y en resumen es un conjunto de servicios de apoyo a la investigación. Los casos de uso se han desarrollado en torno a la gestión de datos de investigación, códigos computacionales, entornos de cómputo, contenidos educativos, proyectos, comunicaciones del grupo de trabajo y la visibilidad web del grupo o proyecto, meidante los servicios *MiLab*.

7.1. Casos de uso

A continuación se describen los casos de uso incorporados a la plataforma MiLab y las actividades realizadas. Se distingue entre *consolidados* para los casos de uso donde se evidencia la

implementación del modelo y *en consolidación* para los casos de uso en etapa de reconocimiento del modelo y de los servicios de la plataforma MiLab. De manera general, en cada caso de uso se requirió de entrenamiento mediante talleres y acompañamiento personalizado al personal sobre las metodologías y herramientas usadas.

7.1.1. Casos de uso consolidados.

7.1.1.1. LA-CoNGA physics. El programa académico *LA-CoNGA physics* es el caso de uso más relevante para la evaluación del modelo propuesto en este trabajo de investigación. *LA-CoNGA physics* es una alianza para crear capacidades en física avanzada a nivel latinoamericano [`https://laconga.redclara.net`]. La implementación del modelo permitió al programa:

- Contar con un plan de gestión de datos desde las primeras etapas del proyecto, disponible en `https://github.com/LA-CoNGA/WP5-Dissemination/blob/master/DataManagementPlan/2021-04-21-LA-CoNGA-DMP.pdf`
- Tener un sistema de gestión de contenidos educativos (LCMS por sus siglas en inglés) para la creación y preservación de los contenidos educativos del programa. El sistema en mención se alimenta de manera colaborativa, se gestiona con control de versiones y se construye y actualiza usando la funcionalidad CI/CD del servicio *G-Lab*. Los contenidos se pueden consultar en `https://laconga.redclara.net/courses/`
- Preservar y gestionar las tareas junto a sus soluciones usando control de versiones [Ver grupos de tareas en `https://gitmilab.redclara.net/laconga`].
- Gestionar las comunicaciones y coordinar sus actividades desde el servicio *chatLab*, esto

ha sido fundamental para el programa debido a las distribución geofigura del personal en el programa.

Este caso de uso permitió también evaluar el modelo en un contexto fuera de la plataforma MiLab, puntualmente en el marco del evento denominado *Hackathon Co-Afina: Datos Abiertos en América Latina 2022* [`https://laconga.redclara.net/hackathon/`]. En este evento estudiantes latinoamericanos resolvieron retos para mejorar la comunicación de la ciencia y la educación usando datos abiertos. Los códigos computacionales creados para analizar los conjuntos de datos se gestionaron desde el servicio `https://github.com/`, las comunicaciones se gestionaron desde el servicio `https://discord.com/` y se coordinó con la red académica ecuatoriana CEDIA [https://cedia.edu.ec] para establecer el entorno de cómputo requerido para las solución de los retos.

7.1.1.2. Grupo Halley (Universidad Industrial de Santander). La implementación del modelo propuesto y la plataforma MiLab permitió al *grupo Halley* de divulgación e investigación en astronomía, gestionar y preservar sus contenidos digitales en ambientes especializados. Algunos de los logros se listan a continuación:

- Se centralizaron las comunicaciones en el servicio *chatLab*.
- Los códigos computacionales se gestionan ahora bajo un sistema de control de versiones en el servicio *G-Lab*. `https://gitmilab.redclara.net/halleyUIS`
- Se preservaron entornos de cómputo mediante el uso de contenedores. `https://hub.docker.com/u/halleyuis`

- Se logró sistematizar el control de actividades de los proyectos con la utilidad de tableros tipo *canvan* del servicio *chatLab* [Ver figura 5].

Figura 5. Captura de pantalla del tablero tipo *canvan* usado por el *grupo Halley* para la gestión de actividades en el marco del programa *LA-CoNGA physics*.

De manera particular destacamos el proyecto *LiMoNet*Peña Rodríguez et al. (2021) en el marco de este caso de uso. Se logró establecer un entorno de trabajo para los análisis de calibración del conjunto de estaciones de medición de esta red. El entorno permite la preservación de códigos computacionales (`https://gitmilab.redclara.net/halleyUIS/limonet`) usando control de versiones. Permite preservar y catalogar de manera automatizada los conjuntos de datos en el sistema *dataLab* (https://dataverse.redclara.net/dataverse/limonet). Realizar los análisis de computo consumiendo los elementos anteriores en el servicio *compLab* (o cualquier servicio *Jupyter Notebook* con acceso a Internet).

7.1.1.3. Proyectos Muografía. Elementos del modelo propuesto se han implementado en el marco de trabajo del grupo de Muografía compuesto por personal del Instituto Tecnoló-

gico Metropolitano de Medellín, la Universidad de Pamplona, el Servicio Geológico Colombiano y la Universidad Industrial de Santander. Se ha logrado gestionar la documentación del proyecto en el servicio *G-Lab* [`https://gitmilab.redclara.net/muografia`] y gestionar las comunicaciones del grupo de trabajo desde el servicio *chatLab*.

7.1.2. Casos de uso en consolidación. En estos casos de uso se han realizado labores de entrenamiento en las temáticas, metodologías y herramientas relacionadas al modelo desarrollado. Se han creado espacios de trabajo en la plataforma MiLab. Las actividades se han centrado principalmente en lograr la centralización de sus comunicaciones usando el servicio *chatLab*. A continuación se listan los casos de uso en etapa de reconocimiento del modelo y de los servicios de la plataforma MiLab.

- Latin American Giant Observatory, LAGO `http://lagoproject.net/`
- Laboratorio de Humanidades Digitales `http://hdlab.space/`
- Laboratorio de Campos y Partículas (Universidad Central de Venezuela).
- Grupo de Investigación en Simbiosis Hospedero- Microorganismo, GISiHM (Universidad de Costa Rica).
- Centro de Investigación en Biología Celular y Molecular (Universidad de Costa Rica).
- Escuela de Física (Universidad Industrial de Santander) `http://fis.uis.edu.co/eisi/`

Tabla 1
Resumen de casos de uso para la validación del modelo

Caso de Uso	Consolidado	Personas involucradas	Actividades y/o Productos
LA-CoNGA physics	Sí	150 Aprox.	- Creación de Plan de Gestión de Datos - Creación de plataforma LCMS - Gestión de comunicaciones (chatLab y Discord) - Gestión de códigos computacionales (G-Lab) - Análisis computacionales en servicio compLab
Grupo Halley UIS	Sí	50 Aprox.	- Gestión de proyectos mediante el uso de tableros chatLab - Gestión de comunicaciones en servicio chatLab - Gestión de códigos computacionales (G-Lab) - Gestión de entornos de cómputo mediante contenedores - Análisis de calibración y preservación de datos del proyecto LiMoNet
Proyectos Muografía	Sí	22	- Gestión documental en servicio G-Lab - Gestión de comunicaciones en servicio chatLab
LAGO	No	100	- Gestión de comunicaciones en servicio chatLab - Interés para gestión de datos en servicio dataLab
HD LAB	No	15	- Interés en realizar análisis de cómputo en temas documentales
Laboratorio de Campos y Partículas	No	10 Aprox.	- Conversaciones para la centralización de comunicaciones y gestión de tareas desde el servicio chatLab
GISiHM	No	20 Aprox	Interés en: - Realizar análisis de cómputo en el servicio compLab - Uso de los servicios chatLab y G-Lab
Centro de Investigación en Biología Celular y Molecular	No	15 Aprox	Interés en: - Realizar análisis de cómputo en el servicio compLab - Uso de los servicios chatLab y G-Lab
Escuela de Física UIS	No	300 Aprox.	- Interés en usar el servicio chatLab para crear un entorno de comunicación de la comunidad académica

7.2. Estadísticas de uso de la plataforma MiLab

La implementación de los anteriores casos de uso sobre la plataforma MiLab tiene a fecha de agosto 18 de 2022, las siguientes estadísticas de uso (tomadas de los paneles de administración).

7.2.1. Servicio *chatLab*.

- Total de usuarios activos: 476
- Promedio de usuarios activos diarios: 45
- Promedio de usuarios activos mensuales: 111
- Total de equipos: 19
- Total de canales (públicos y privados): 310
- Total de mensajes: 116258

Figura 6. Estadísticas generales de uso del servicio *chatLab*.

En la figura 7 se ilustra el comportamiento del envío de mensajes diario en la plataforma para el periodo 2022-07-17 a 2022-08-16.

7.2.2. Servicio *G-Lab*.

- Total de usuarios: 405 [Ver figura 8]

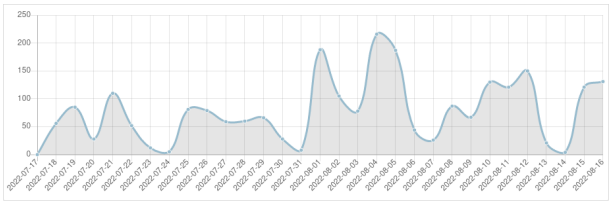

Figura 7. Estadísticas de envío de mensajes en el servicio *chatLab* para el periodo 2022-07-17 a 2022-08-16.

- Total de proyectos: 1388 [Ver figura 9]
- Total de grupos: 55
- Total de pipelines(acciones CI/CD): 2025 [Ver figura 9]
- Total de Issues: 450 [Ver figura 9]
- Total de Issues: 310 [Ver figura 9]

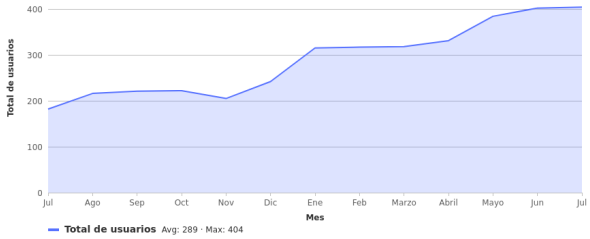

Figura 8. Comportamiento del total de usuarios del servicio *G-Lab* en el periodo julio-2021 a julio-2022.

7.2.3. Servicio *dataLab*. El servicio *dataLab* ha sido el de menor uso en el marco de los casos de uso trabajados, la estadísticas de uso se limitan a los conjuntos de datos gestionados desde el proyecto *LiMoNet*. Específicamente, 187 datasets publicados de los cuales 15 se

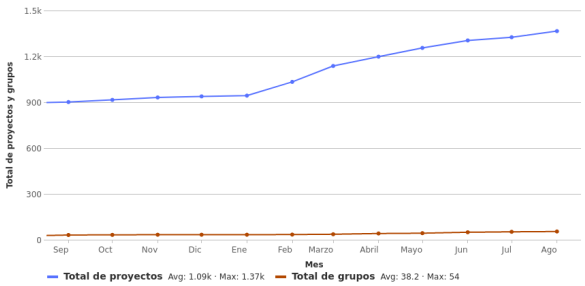

Figura 9. Comportamiento del total de proyectos del servicio *G-Lab* en el periodo septiembre-2021 a agosto-2022.

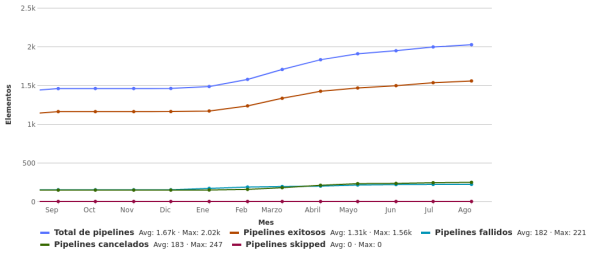

Figura 10. Comportamiento del total de pipelines (acciones CI/CD) del servicio *G-Lab* en el periodo septiembre-2021 a agosto-2022.

encuentran en acceso abierto y 172 de acceso privado al grupo de trabajo del proyecto [Ver figura 12].

8. Conclusiones

Desde el trabajo realizado en este trabajo de investigación para aportar a la solución de la crisis en la reproducibilidad de la ciencia , es posible enunciar las siguientes conclusiones.

- Se logró una caracterización del escenario de investigación con los elementos esenciales para la reproducibilidad.
- El modelo desarrollado permitió mejorar el grado de reproducibilidad del trabajo investigativo en los casos de uso implementados.

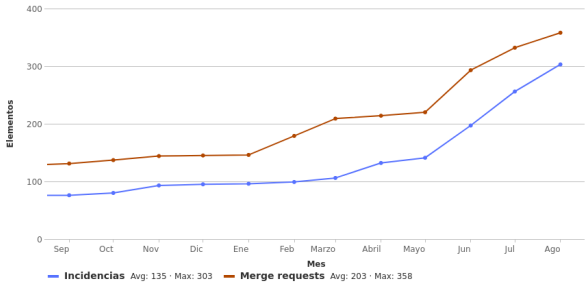

Figura 11. Comportamiento del total de incidencias y Merge requests del servicio *G-Lab* en el periodo septiembre-2021 a agosto-2022.

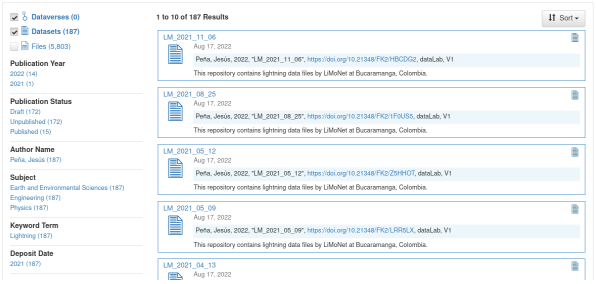

Figura 12. Dataset publicados en el servicio *dataLab*.

- Se logró crear entornos de trabajo en los que de manera centralizada, se preservaron los elementos esenciales del proceso de investigación.

- El modelo se reconoce como una oportunidad de negocio sostenible y escalable en escenarios de investigación.

Algunos comentarios sobre el trabajo realizado y el contexto en que se realizó.

- El escenario de producción de conocimiento ha cambiado considerablemente y la Ciencia Abierta ha sido uno de los grandes motores de cambio. Los requisitos para acceder a fondos de financiación, las políticas de Ciencia Abierta y las directivas de entidades gubernamentales, están definiendo las nuevas formas de crear y compartir conocimiento. Partiendo de

la base de que los productos generados con financiamientos públicos son patrimonio de la humanidad y deben estar accesibles y disponibles sin restricción, el compartir se hace aún más relevante. Sin lugar a dudas este nuevo escenario tiene grandes batallas por librar y estamos en el momento adecuado para contribuir. Desde vencer el miedo en el personal de investigación a compartir sus datos, hasta las disputas legales con las grandes editoriales y/o oficinas de patentes.

- La reproducibilidad de experimentos científicos es la base de la ciencia misma. Ofrecer todos los elementos que permitan la reproducción de un un experimento debe ser un objetivo a cumplir. Así como lo enuncia *The Turing Way*, la reproducibilidad "debe ser tan fácil que lo difícil sea no hacerlo".

- El escenario actual es prometedor para la reproducibilidad de la ciencia, especialmente en la solución de su crisis. Los avances metodológicos y tecnológicos en los procesos de investigación se han dado a pasos agigantados en la última década. Entre otros, la pandemia COVID 19 ha servido como catalizador de muchos de estos avances.

- Finalmente, destacamos la importancia de la formación del personal de investigación en las nuevas formas de crear conocimiento. Tendremos éxito en la solución a la crisis de reproducibilidad o la implementación de las nuevas políticas de ciencia abierta, si como comunidad académica o incluso como sociedad, adquirimos las habilidades requeridas.

Como parte de las actividades de divulgación del conocimiento se generaron los siguientes productos.

- Artículo en revista *UIS Ingenierías*. Título del artículo: Academia, datos y reproducibilidad de la ciencia. Vol. 19, n.° 4, pp.315-324, 2020 (Autores: Martínez-Méndez, A., Nuñez, L. A.) `https://doi.org/10.18273/revuin.v19n4-2020026`

- Ponencia oral en el *Congreso Iberoamericano de Ciencia Abierta* (2022, del 23 al 24 de noviembre. Online) [Título de la ponencia: Modelo de Gestión de Recursos Computacionales para Asistir la Reproducibilidad de Experimentos Científicos]

- Ponencia oral en el evento *OpenCon Latam 2019* (Bogotá, Colombia, 26 - 28 de septiembre de 2019) [Título de la ponencia: Gestión de Datos en la Colaboración LAGO, Estrategias y Reflexiones]

- Ponencia oral en el *13th LAGO Workshop* (Tucumán, Argentina, 21 - 26 de febrero, 2022) [Título de la ponencia: MiLab: Plataforma como servicio en la nube para fomentar la reproducibilidad en grupos de investigación]

- Poster presentado en *The Latin America High Performance Computing Conference, CARLA* (Porto Alegre, Brazil, septiembre 26-30, 2022) [Título del trabajo: MiLab: A platform for reproducible research]

- Poster en el *Congreso Colombiano de Astronomía, CoCoA* (Tunja, Colombia, 21 - 23 de septiembre de 2022) [Título del trabajo: MiLab: Plataforma para la Reproducibilidad de la Ciencia]

- Taller en el *Congreso de Tecnologías de la Información y Comunicación TICEC 2021*, Ecua-

dor. (Noviembre 26 - 29, 2021) [Título del taller: Supporting Open Research with MiLAB]

Adicionalmente, durante el desarrollo de este trabajo tuve la oportunidad de participar del programa de embajadores de Ciencia Abierta *Open Life Science, OLS*. Inicialmente como participante en la tercer cohorte con el proyecto LA-CoNGA physics `https://openlifesci.org/ols-3/projects-participants/#mxrtinez`. Después como tutor en la cuarta y quinta cohorte `https://openlifesci.org/ols-4#mentors` `https://openlifesci.org/ols-5#mentors`

Referencias Bibliográficas

(2017). LEARN Toolkit of Best Practice for Research Data Management. Technical report, Learn.

Ailamaki, A., Kantere, V., and Dash, D. (2010). Managing scientific data. *Communications of the ACM*, 53(6):68.

Allison, D. B., Brown, A. W., George, B. J., and Kaiser, K. A. (2016). Reproducibility: A tragedy of errors. *Nature*, 530(7588):27–29.

Andrew, A. (2018). Essay: The Experiments Are Fascinating. But Nobody Can Repeat Them.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454.

Ballhausen, M. (2019). Free and open source software licenses explained. *Computer*, 52(6):82–86.

Barba, L. A. (2016). The hard road to reproducibility. *Science*, 354(6308):142–142.

Begley, C. G. and Ioannidis, J. P. (2015). Reproducibility in Science: Improving the Standard for Basic and Preclinical Research. *Circulation Research*, 116(1):116–126.

Brammer, G. R., Crosby, R. W., Matthews, S. J., and Williams, T. L. (2011). Paper Mâché: Creating Dynamic Reproducible Science. *Procedia Computer Science*, 4:658–667.

Buck, S. (2015). Solving reproducibility. *Science*, 348(6242):1403–1403.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., et al. (2018). Evaluating the

replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644.

Chalstrey, E. (2021). Developing and Publishing Code for Trusted Research Environments: Best Practices and Ways of Working. *arXiv:2111.06301 [cs]*.

Chen, X., Dallmeier-Tiessen, S., Dasler, R., Feger, S., Fokianos, P., et al. (2019). Open is not enough. *Nature Physics*, 15(2):113–119.

Committee, U. R. N. S. (2021). From grassroots to global: A blueprint for building a reproducibility network. *PLOS Biology*, 19(11):e3001461.

Community, T. T. W., Arnold, B., Bowler, L., Gibson, S., Herterich, P., et al. (2019). The Turing Way: A Handbook for Reproducible Data Science.

Community, T. T. W. and Scriberia (2020). Illustrations from the Turing Way book dashes.

Dcc (2013). Checklist for a Data Management Plan, v4.0.

de Waard, A., Cousijn, H., and Aalbersberg, I. J. (2015). 10 aspects of highly effective research data. *Elsevier Connect*.

Delevante, R. (2019). 5 trends in research data management. *Elsevier Connect*.

Dirnagl, U. (2020). Institutions can retool to make research more rigorous. *Nature*, pages d41586–020–02905–1.

DMA, A. B. (2019). Reproducibility in research: taming a “complex beast”. *Elsevier Connect*.

Educational, U. N., UNESCO, , , , et al. (2021). Recomendación de la UNESCO sobre la Ciencia Abierta.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115(11):2628–2631.

Feger, S. S., Dallmeier-Tiessen, S., Schmidt, A., and Woźniak, P. W. (2019). Designing for Reproducibility: A Qualitative Study of Challenges and Opportunities in High Energy Physics. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, pages 1–14, Glasgow, Scotland Uk. ACM Press.

Felter, W., Ferreira, A., Rajamony, R., and Rubio, J. (2015). An updated performance comparison of virtual machines and Linux containers. In *2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 171–172, Philadelphia, PA, USA. Ieee.

Ferreira, P., Baron, T., Bossy, C., Pugh, M., Resco, A., et al. (2012). Indico: A Collaboration Hub. In *Journal of Physics: Conference Series*, volume 396, page 062006. IOP Publishing.

Florez Vargas, O. (2016). *Development of strategies for assessing reporting in biomedical research: moving toward enhancing reproducibility*. PhD Thesis, The University of Manchester, Manchester.

for Science, C. and Technology Studies, CWTS, B. (2017). Open data: The researcher perspective.

Goodman, A., Peek, J., Accomazzi, A., Beaumont, C., Borgman, C. L., et al. (2017). The "Paper.ºf the Future. Technical report, Authorea, Inc.

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12–341ps12.

Gregg, W., Erdmann, C., Paglione, L., Schneider, J., and Dean, C. (2019). A literature review of scholarly communications metadata. *Research Ideas and Outcomes*, 5:e38698.

Haak, W. (2019). 4 principles for unlocking the full potential of research data. *Elsevier Connect*.

Ioannidis, J. P. A. (2014). How to Make More Published Research True. *PLoS Medicine*, 11(10):e1001747.

Ivie, P. and Thain, D. (2016). PRUNE: A preserving run environment for reproducible scientific computing. In *2016 IEEE 12th International Conference on e-Science (e-Science)*, pages 61–70, Baltimore, MD, USA. Ieee.

Jupyter, P., Bussonnier, M., Forde, J., Freeman, J., Granger, B., et al. (2018). Binder 2.0 - Reproducible, interactive, sharable environments for science at scale. pages 113–120, Austin, Texas.

Kikkenborg, E. (2019). Up2U-Up to University-Bridging the digital gap between schools and universities. *Impact*, 2019(1):12–13.

Koers, H. (2016). How do we make it easy and rewarding for researchers to share their data? A publisher's perspective. *Journal of Clinical Epidemiology*, 70:261–263.

Leisch, F., Eugster, M., and Hothorn, T. (2011). Executable Papers for the R Community: The R2 Platform for Reproducible Research. *Procedia Computer Science*, 4:618–626.

Macdonald, S. and Macneil, R. (2015). Service Integration to Enhance Research Data Management: RSpace Electronic Laboratory Notebook Case Study. *International Journal of Digital Curation*, 10(1):163–172.

Manghi, P., Manola, N., Horstmann, W., and Peters, D. (2010). An infrastructure for managing EC funded research output - The OpenAIRE Project. *The Grey Journal (TGJ) : An International Journal on Grey Literature*, 6(1):31–40.

Markowetz, F. (2015). Five selfish reasons to work reproducibly. *Genome Biology*, 16(1):274.

Mondelli, M. L., Peterson, A. T., and Gadelha Jr, L. M. R. (2019). Exploring Reproducibility and FAIR Principles in Data Science Using Ecological Niche Modeling as a Case Study. *arXiv:1909.00271 [cs]*.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):0021.

Nature (2021). Data Repository Guidance. *Scientific Data*.

Nelson, A. (2022). Memorandum for the heads of executive departments and agencies. policy guidance to federal agencies with research and development expenditures on updating their public access policie.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., et al. (2015). Promoting an open research culture. *Science*, 348(6242):1422–1425.

Pain, E. (2018). Meet Octopus, a new vision for scientific publishing. *Science*.

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance*, 12(3):30–32.

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, 334(6060):1226–1227.

Perkel, J. M. (2020). Challenge to scientists: does your ten-year-old code still run? *Nature*, 584(7822):656–658.

Peña Rodríguez, J., Salgado-Meza, P., Flórez-Villegas, L., Peña-Rodríguez, J., and Núñez, L. A. (2021). A lightning detection system for studying transient phenomena in cosmic ray observatories. *PoS*, Icrc2021:253.

Piater, J. (2005). A Guide to Coding Style.

Piparo, D., Tejedor, E., Mato, P., Mascetti, L., Moscicki, J., et al. (2018). SWAN: A service for interactive analysis in the cloud. *Future Generation Computer Systems*, 78:1071–1078.

Poldrack, R. A., Gorgolewski, K. J., and Varoquaux, G. (2019). Computational and Informatic Advances for Reproducible Data Analysis in Neuroimaging. *Annual Review of Biomedical Data Science*, 2(1):119–138.

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., et al. (2016). Enhancing reproducibility for computational methods. *Science*, 354(6317):1240–1241.

Teytelman, L., Stoliartchouk, A., Kindler, L., and Hurwitz, B. L. (2016). Protocols.io: Virtual Communities for Protocol Development and Discussion. *PLOS Biology*, 14(8):e1002538.

Thomas, K., Benjamin, R.-K., Fernando, P., Brian, G., Matthias, B., et al. (2016). Jupyter Notebooks; a publishing format for reproducible computational workflows. *Stand Alone*, pages 87–90.

Trisovic, A. (2018). *Data preservation and reproducibility at the LHCb experiment at CERN*. PhD Thesis.

Van Atteveldt, W., Strycharz, J., Trilling, D., and Welbers, K. (2019). Toward Open Computational Communication Science: A Practical Road Map for Reusable Data and Code. *International Journal of Communication*.

van der Graaf, M. and Waaijers, L. (2011). A Surfboard for Riding the Wave. Towards a four country action programme on research data. A Knowledge Exchange Report. Technical report.

Vrieze, J. (2021). Landmark research integrity survey finds questionable practices are surprisingly common. *Science*.

Whitaker, K. (2018). Barriers to reproducible research (and how to overcome them). page 3070929 Bytes.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Willems, L. (2019). 6 insights from leading universities on managing research data effectively. *Elsevier Connect*.

Wilson, G. (2016). Software Carpentry: lessons learned. *F1000Research*, 3:62.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., et al. (2017). Good enough practices in scientific computing. *PLOS Computational Biology*, 13(6):e1005510.

Xie, Y. and Allaire, J. (2012). New Tools for Reproducible Research with R. Nashville, Tennessee, USA.

Yang, Y., Youyou, W., and Uzzi, B. (2020). Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(20):10762–10768.

Apéndices

Apéndice A. Herramientas para la reproducibilidad

Las herramientas tecnológicas actuales ofrecen mejores escenarios para generar ambientes que faciliten la gestión de grupos de investigación y la reproducibilidad de sus actividades. En este apéndice mostramos algunas de las herramientas encontradas en el marco de este trabajo de investigación para apoyar los procesos de investigación en cada una de sus etapas. Desde la captura de los conjuntos de datos, hasta la publicación de los resultados [Ver imágen 13].

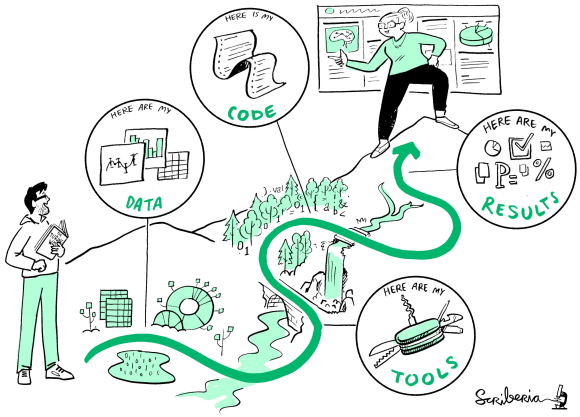

Figura 13. Ruta para obtener publicaciones con alto grado de reproducibilidad por *The Turing Way*. Adaptado de Community and Scriberia (2020).

Repositorios de datos. Los repositorios de datos son la mejor alternativa para preservar y catalogar datos de investigación. En el mercado existe una gran variedad de estos, algunos creados específicamente por disciplina o proyecto como el es caso del Global Biodiversity Information Facility (GBIF) para datos de biodiversidad y su proyecto de ciencia ciudadana `https://www.gbif.org/`. Otros de propósito general y sobre los cuales nos enfocamos en este

trabajo de investigación por su versatilidad. Además de la preservación y tal como se menciona en Nature (2021) o en las recomendaciones por de-Ward en de Waard et al. (2015), los repositorios de datos deben permitir mínimamente el acceso a los conjuntos de datos, su búsqueda, su reusabilidad y el reconocimiento a sus autores(as), esto se resume de gran manera en el cumplimiento de los principios FAIR Wilkinson et al. (2016). A continuación mencionamos 6 de los sistemas de repositorio de datos más usados en la actualidad con enfoque en acceso abierto.

- Figshare. Disponible en `https://figshare.com/`
- Mendeley Data. Disponible en `https://data.mendeley.com/`
- Dryad Digital Repository. Disponible en `https://datadryad.org/stash`
- Harvard Dataverse. Disponible en `https://dataverse.harvard.edu/` (usado en la plataforma *MiLab* producto de este trabajo de investigación).
- Open Science Framework. Disponible en `https://osf.io/`
- Zenodo. Disponible en `https://zenodo.org/`

Sistemas de control de versiones. Más allá de los muy diversos lenguajes de programación para la creación de códigos computacionales y las buenas prácticas de programación, la trazabilidad de los cambios realizados y los registros de las personas involucradas son elementos esenciales para favorecer la reproducibilidad. Destacamos los *Sistemas de Control de Versiones* (CVS por sus siglas en inglés) como la herramienta más adecuada para obtener estas dos características. Estos sistemas nacieron con el desarrollo de software y en resumen, nos permiten registrar

cualquier cambio realizado e ir entre cada una de las versiones registradas, esto acompañado de información sobre las personas involucradas. En la figura 14 se ilustra de manera gráfica con un ejemplo la trazabilidad lograda usando un *CVS* sobre un proyecto, las ramas en ilustran la posibilidad de probar nuevas funcionalidades y corregir errores simultáneamente.

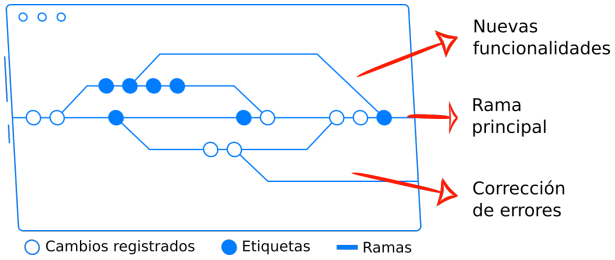

Figura 14. Representación tipo árbol del historial de cambios en un proyecto gestionado con control de versiones. Adaptado de *storyset on Freepik*.

Los *CVS* permiten también la preservación de los códigos computacionales y facilitan de gran manera el trabajo colaborativo. *GitHub* [`https://github.com/`] y *GitLab* [`https://gitlab.com/`] como servicios basados en un *CVS*, específicamente *git* [`https://git-scm.com/`], ofrecen una plataforma de servicios para la creación de software de manera colaborativa, el desarrollo continuo, además de contar con funcionalidades para la gestión de proyectos.

Contenerización y virtualización. El software y el hardware determinan el entorno de un análisis computacional, en esta sección nos enfocamos en las alternativas para encapsular el software mediante virtualización y/o contenerización. La virtualización intenta emular una máquina física completa, mientras que la contenerización busca aislar una aplicación en el host. Con el objetivo de preservar el entorno y facilitar su portabilidad o despliegue, su uso se puede dar por separado o en conjunto dependiendo de las características del entorno. De manera similar

ocurre con el rendimiento, este depende de factores como las herramientas usadas y la arquitectura, sin embargo en términos generales la contenerización ofrece unos mejores resultados Felter et al. (2015). La figura 15 ilustra de manera resumida la arquitectura de estas dos tecnologías.

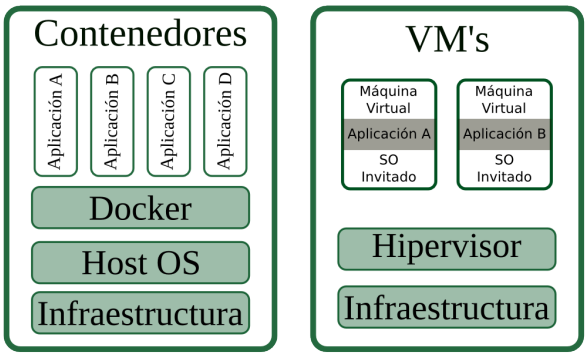

Figura 15. Arquitectura de contenerización (izquierda) vs. virtualización (derecha).

Dentro de los sistemas más populares para la gestión de contenedores encontramos: Docker, Podman, Kubernetes, LXC, Microsoft Azure Container Registry, Containerd y Vagrant. Para las máquinas virtuales algunos de los hipervisores más populares son: Virtual Box, VMware, Oracle VM VirtualBox, Hyper-V Manager y QEMU.

En términos de implementación de estas tecnologías, especialmente de la contenerización, en el marco de este trabajo encontramos la herramienta *binder*. Esta permite de manera ágil crear y desplegar entornos software de cómputo en la nube mediante contenedores Jupyter et al. (2018) y repositorios en sistemas *git*, el proceso y arquitectura se ilustra en la figura 16.

Herramientas de comunicación. La gestión de las comunicaciones en entornos de investigación o laborales requiere de dos elementos principales, la centralización y preservación. En este sentido herramientas de mensajería instantánea (Whatsapp, Telegram, Messenger, etc.)

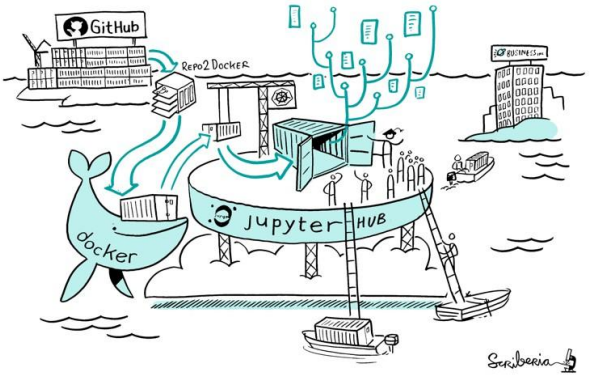

Figura 16. Arquitectura del servicio `https://mybinder.org/`. Tomado de *The Turing Way Scriberia* Community and Scriberia (2020).

o redes sociales, aunque ampliamente utilizadas, no son la mejor alternativa. La volatilidad de la información aumenta y su preservación se hace difícil al estar en herramientas de terceros. Presentamos a continuación algunas de las herramientas revisadas en el marco de este trabajo de investigación donde además de la preservación y centralización, se revisó la capacidad de realizar búsquedas, generar estadísticas y permitir la integración con otros sistemas.

- Slack. Disponible en `https://slack.com/`
- Discord. Disponible en `https://discord.com/`
- RocketChat. Disponible en `https://rocket.chat/`
- Mattermost. Disponible en `https://mattermost.com/` (usada en la plataforma *MiLab* producto de este trabajo de investigación).

ELN - Libretas de Laboratorio/Campo. Las libretas de laboratorio o de campo encuentran en los Cuadernos Electrónicos de Laboratorio, ELN (por sus siglas en inglés, Electro-

nic Lab Notebook) la evolución de las notas a mano. El uso de ELNs facilita la preservación de información y su interpretación. Dos herramientas disponibles en el mercado con funcionalidades de ELN son:

- Protocols.io Teytelman et al. (2016). Disponible en `https://www.protocols.io/` esta herramienta se enfoca en la preservación de los métodos usados en el proceso de investigación. *Protocols.io* permite entre otros, documentar los materiales y equipos usados, las personas involucradas y marcas temporales de cada uno de los procesos.

- RSpace Macdonald and Macneil (2015). Disponible en `https://www.researchspace.com/` es una herramienta integradora de servicios para apoyar la reproducibilidad y la implementación de los principios FAIR. La figura 17 ilustra el conjunto de servicios integrados en esta herramienta, desde las etapas de preparación, la investigación activa, hasta la preservación y reuso de los elementos generados en los experimentos.

Notebooks Jupyter. Los *Notebooks Jupyter* Thomas et al. (2016) lograron integrar en una sola interfaz códigos computacionales, documentación y resultados favoreciendo la reproducibilidad de los análisis. Desde un enfoque interactivo esta herramienta permite desarrollar códigos computacionales en docenas de lenguajes de programación, además de tener un enfoque en el desarrollo de código abierto y estándares abiertos. La arquitectura de esta herramienta tiene 3 elementos principales, el lado del cliente dónde desde una interfaz web se interactúa con el Notebook, el servidor Jupyter donde se gestionan los Notebooks (Archivos *json*) y los Kernels donde se realiza el cómputo [Ver figura 18]. Estos últimos pueden estar en el equipo local o pueden estar

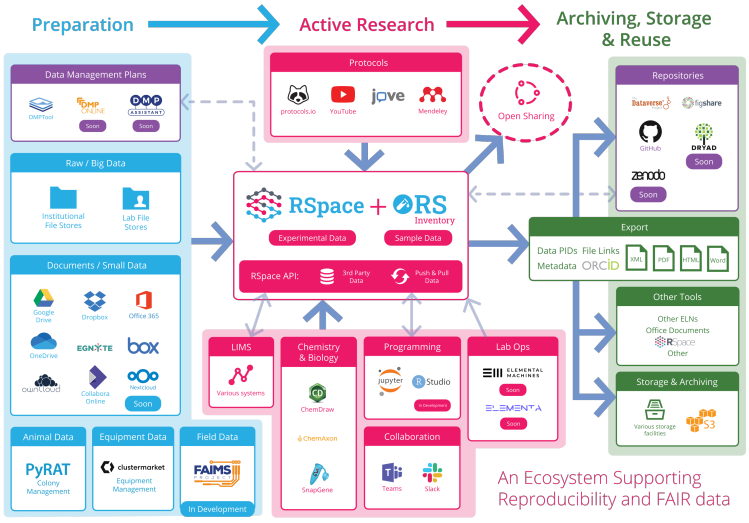

Figura 17. Arquitectura del ecosistema RSpace para el soporte a la reproducibilidad y los datos FAIR. Tomado de `https://www.researchspace.com/`

en equipo remotos permitiendo así el trabajo sobre servicios cómputo más complejos.

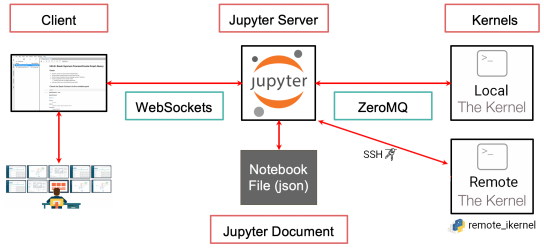

Figura 18. Arquitectura de los Notebooks Jupyter. Adaptado de `https://securitydatasets.com/consume/jupyter-notebooks.html`

SWAN. *SWAN* es una plataforma que integra algunas de las herramientas mencionadas anteriormente con la intención de ofrecer un entorno interactivo de análisis en la nube Piparo et al. (2018). Desarrollada por el CERN esta plataforma integra servicios de cómputo, contenerización, repositorios de datos y servicios de almacenamiento para apoyar el trabajo de investigadores(as). Es una herramienta de grandes capacidades, sin embargo su uso está restringido a personal del *CERN*. En la figura 19 se exponen los pilares de esta plataforma, cómputo, software

y almacenamiento, además de las herramientas usadas en cada uno de estos.

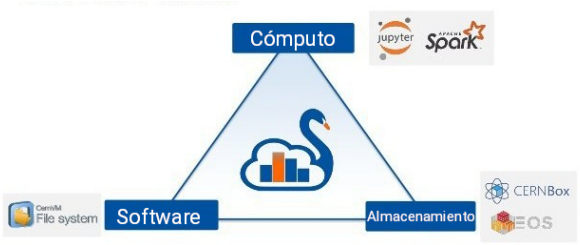

Figura 19. Servicios de la plataforma para análisis interactivos en la nube *SWAN*. Adaptado de Piparo et al. (2018).

Apéndice B. Plataforma MiLab

Presentamos la plataforma MiLab [Ver figura 20], una plataforma integradora para el apoyo a la investigación y el control de proyectos, creada sobre herramientas abiertas para garantizar la gobernanza de la información, con la reproducibilidad como eje principal. MiLab desde una estructura modular busca preservar, facilitar el acceso y dar trazabilidad al mayor número de elementos en los procesos de investigación, principalmente de:

- Datos de investigación
- Códigos computacionales
- Comunicaciones
- Cálculo computacional
- Laboratorios de investigación
- Visibilidad web y documentación

La plataforma MiLab la compone un conjunto de 6 servicios [Ver figura 20] los cuales se integran mediante el uso de API's, plugins y un servicio de autenticación federado. Los servicios MiLab se incorporan a los flujos de trabajo en escenarios de investigación, desde la fuente de los datos, hasta la publicación de los resultados [Ver figura 21]. La gestión de las comunicaciones y proyectos del grupo o centro de investigación son elementos transversales. A excepción del servicio *compLab*, los demás servicios se ofrecen desde un entorno compartido que permita la interacción entre la comunidad. A continuación se describen los servicios MiLab.

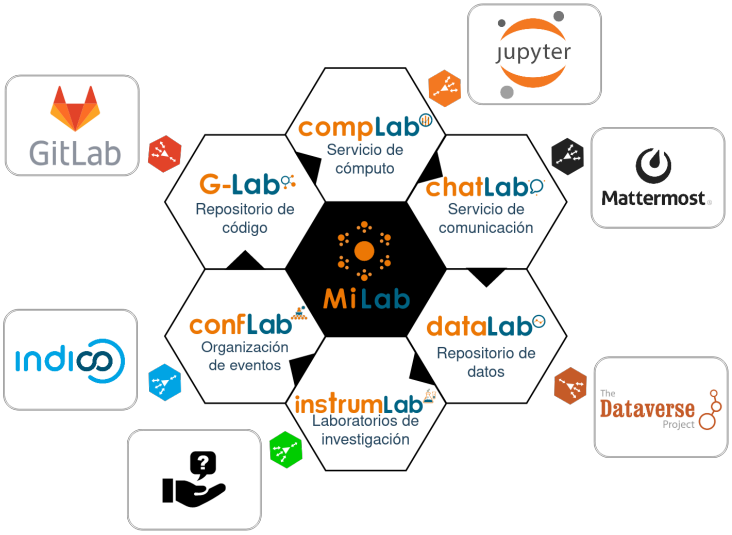

Figura 20. Plataforma MiLab para el apoyo a la investigación y el control de proyectos.

dataLab - Gestión de Datos de investigación. *dataLab* es un sistema de repositorio para la preservación y difusión de datos de investigación, específicamente usa el software Dataverse [`https://dataverse.org/`], desarrollado desde el *Institute for Quantitative Social Science (IQSS)* con el apoyo de Harvard University. *dataLab* permite catalogar y preservar en un lugar seguro y de fácil acceso los datos usados en los análisis de un proyecto de investigación. A continuación algunas de las características y funcionalidades del servicio.

- Permite gestionar los conjuntos de datos desde su adquisición/recolección/medición/creación hasta su preservación.
- Los conjuntos de datos pueden tener periodos de embargo según las políticas de datos del proyecto.
- Los conjuntos de datos pueden ser catalogados con esquemas de metadatos estandarizados.

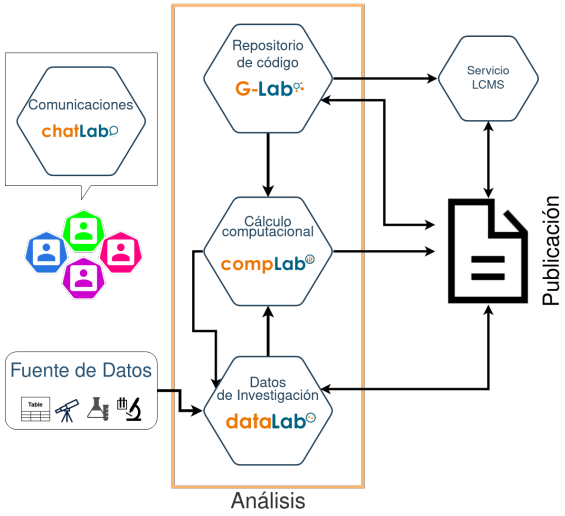

Figura 21. Flujo de trabajo en entornos de investigación mediante los servicios *MiLab*.

- Los derechos de uso de los conjuntos de datos se presentan claramente desde la interfaz web.
- El API Dataverse facilita la ingesta y/o uso de grandes cantidades de datos desde otros sistemas.
- Identificadores digitales (DOI) son asignados a los conjuntos de datos para garantizar, entre otros, la autoría de estos.

Resumiendo, el uso del servicio *dataLab* y siguiendo las recomendaciones del capítulo 6 de este documento, facilita el cumplimiento de los principios **FAIR**.

G-Lab - Gestión de Códigos Computacionales. *G-Lab* se enfoca en la gestión de códigos computacionales mediante un sistema de control de versiones. Utiliza el software *GitLab* para garantizar la trazabilidad de los distintos códigos computacionales creados o usados en el proyecto de investigación. Usando *git* este servicio permite preservar el historial de cambios junto a la información de las personas involucradas en un ambiente web.

Posee también otras funcionalidades para la gestión de proyectos, el desarrollo de software y el trabajo colaborativo. Mediante el registro de incidencias (Issues), tareas, etiquetas, tableros (Boards) e hitos (Milestones) permite la gestión de proyectos, especialmente usando metodologías ágiles. Finalmente, usando las capacidades de integración continua y entrega continua, CI/CD (por sus siglas en inglés) se pueden automatizar procesos como el compilado, testeo o despliegue de aplicaciones.

compLab - Cálculo computacional. El servicio *compLab* provee de las capacidades de cómputo necesarias para ejecutar los análisis de cómputo de cada proyecto de investigación desde un ambiente web. Este servicio se ofrece de acuerdo a las capacidades de cómputo requeridas por el grupo o centro de investigación. Mediante contenedores se hacen despliegues aislados de servicios *JupyterHUB* con la interfaz *JupyterLab* defecto. Para capacidades de cómputo de mayores prestaciones se hace uso de kernels remotos.

chatLab - Gestión de Comunicaciones. *chatLab* es un entorno especializado de comunicación, permite centralizar las comunicaciones asíncronas de manera ordenada mediante el uso de canales (públicos o privados) por temática. Permite también preservar la información para así evitar que esta desaparezca de la historia del grupo o proyecto. Se usa el software *Mattermost* para garantizar estas dos funcionalidades. Otras funcionalidades de este servicio permiten gestionar las tareas de los proyectos con el uso de tableros, automatizar procesos con listas de chequeo o integrar servicios de videollamada para coordinar encuentros virtuales.

instrumLab - Gestión de Laboratorios. El servicio *instrumLab* se encuentra en etapa de diseño y actualmente se tienen planteadas las siguientes funcionalidades básicas.

- Uso de ELN para la gestión de libretas de laboratorio o de campo.
- Documentación y preservación de los protocolos usados en los experimentos
- Gestión y trabajo en laboratorios remotos.

confLab - Gestión de Eventos. *confLab* es un servicio para la gestión, de inicio a fin, de eventos científicos, conferencias, workshops o cualquier evento de divulgación y de cualquier complejidad. Actualmente se encuentra en etapa de despliegue y usará el software *indico*Ferreira et al. (2012). En términos de funcionalidades este servicio permitirá gestionar las inscripciones, las salas (físicas o virtuales), el calendario y el material audiovisual, además de facilitar la divulgación del evento mediante una página web.

Integración de Servicios. La integración de los servicios MiLab es fundamental para la percepción de esta como una plataforma y no de un conjunto de servicios aislados. La adopción de metodologías y tecnologías para la reproducibilidad será más eficaz si se logran integrar a las actividades del quehacer investigativo. Dicha integración se entiende como un elemento transversal y se logra en diferentes niveles, desde la autenticación de usuarios hasta el flujo de información entre los servicios. Distinguimos los siguientes niveles de integración.

8.0.0.1. Autenticación. La autenticación es un elemento de especial atención para la plataforma MiLab, es necesario el uso de estándares y protocolos de autenticación que permitan integrar diversos componentes y la agrupación de perfiles para la asignación diferenciada tanto de recursos como de permisos. Específicamente se utilizó un servicio de inicio de sesión simple (SSO por sus siglas en inglés, Single Sign On) basado en el estándar SAML (Security Assertion Markup

Language).

La arquitectura de este sistema SSO la componen varios elementos los cuales se ilustran en la figura 22. La aplicación **SimpleSAMLphp** se encarga de manejar la autenticación en los servicios de acuerdo a la información de usuario registrada en un directorio LDAP (Lightweight Directory Access Protocol). En los servicios *compLab* y *chatLab* debido a limitaciones técnicas, la autenticación se realiza mediante el servicio *G-Lab* usando el protocolo **OAuth**.

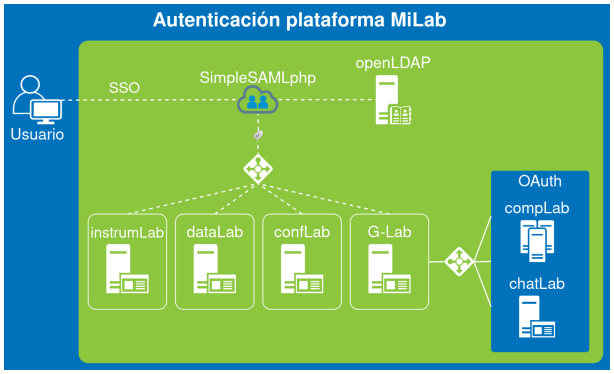

Figura 22. Arquitectura de autenticación federada para los servicios *MiLab*.

8.0.0.2. Autorización. El acceso a los servicios y la privacidad de la información lo gestiona la capa de autorización, cada servicio cuenta con la capacidad de limitar a los usuarios el acceso a recursos, API's o funcionalidades. A continuación característicasicas más relevantes de autorización en los servicios MiLab.

- **chatLab**: Los canales de comunicación pueden ser públicos o privados; Las conversaciones 1-1 o 1-N son privadas; existe al menos un usuario administrador con la capacidad de gestionar usuarios y los espacio de trabajo.

- **compLab**: Este es el único servicio MiLab con despliegues aislados para el grupo o centro

de investigación y posee las características de autorización de un sistema operativo UNIX. Así, la autorización se divide en la capacidad de *Lectura, Escritura y Ejecución* por parte de *Usuario, Grupos y Otros (diferente usuario y/o grupo)* de los archivos o directorios en el espacio de trabajo. Más información en `https://es.wikipedia.org/wiki/Permisos_de_acceso_a_archivos`

- **G-Lab**: La visibilidad de los proyectos y grupos puede ser definida como pública (visibles sin restricción alguna), interna (visibles a los usuarios autenticados en el servicio) o privada (visible solo a miembros del proyecto); A nivel de usuario se tienen 5 roles con diferentes habilidades (Invitado, Reportero, Desarrollador, Mantenedor y Propietario) las cuales se describen en `https://gitmilab.redclara.net/help/user/permissions`.

- **dataLab**: La organización de la información se da en Dataverses y Datasets, con la característica de que se pueden tener Dataverses dentro de cualquier Dataverse. En las figuras 23 y 24 se ilustra respectivamente la organización de Dataverses y los elementos en Datasets; Los roles de usuario definen los permisos para la gestión de Dataverses, Datasets y Archivos, estos se describen de manera detallada en `https://guides.dataverse.org/en/latest/user/dataset-management.html?highlight=roles`

8.0.0.3. API's y Plugins. El uso de API's y/o Plugins permite la interacción entre servicios y ofrece funcionalidades extra a nivel de usuario. Su uso es de gran importancia para lograr integrar los servicios, sin embargo la complejidad tecnológica aumenta con su implementación. Especialmente en términos de mantenimiento y adecuación a los distintos protocolos y

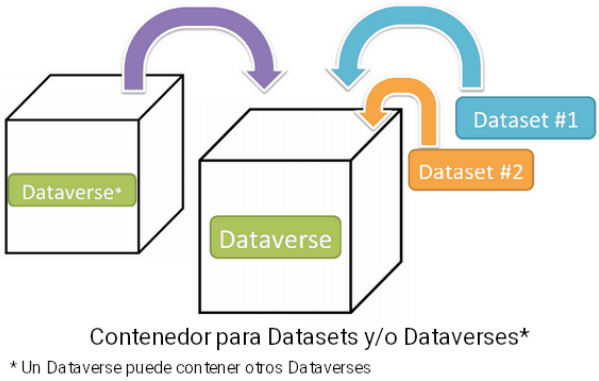

Figura 23. Diagrama de organización de Dataverses en el servicio *dataLab*. Adaptado de `https://guides.dataverse.org/en/4.5/user/dataverse-management.html`

Figura 24. Diagrama de organización de Datasets en el servicio *dataLab*. Adaptado de `https://guides.dataverse.org/en/latest/user/dataset-management.html`

estándares. A continuación se describe el uso de API's y Plugins entre los servicios MiLab.

- Uso del plugin *GitLab Plugin* para enlazar las notificaciones del servicio *G-Lab* al servicio *chatLab*. Más información de este plugin en `https://mattermost.gitbook.io/plugin-gitlab/`

- Uso del API dataverse para gestionar y consumir los conjuntos de datos del servicio *dataLab* usando códigos computacionales en el servicio *compLab*. Más información del API dataver-

se y sus funcionalidades en `https://guides.dataverse.org/en/latest/api/intro.html`.

- Uso del plugin *OAuthenticator* para permitir la autenticación en el servicio *compLab* desde el servicio *G-Lab* usando el protocolo OAuth. Más información de este plugin en `https://oauthenticator.readthedocs.io/en/latest/`

- Uso del plugin *jupyterlab-git* para gestionar repositorios git del servicio *G-Lab* (o de cualquier otro servicio similar) de manera gráfica desde la interfáz web *JupyterLab*. Más información de este plugin en `https://pypi.org/project/jupyterlab-git/`

Especificaciones de Cómputo Actuales. La plataforma MiLab se ejecuta actualmente sobre un conjunto de servidores virtuales de la Red Académica Ecuatoriana CEDIA, las especificaciones se pueden ver en la tabla 2.

Tabla 2
Especificaciones de cómputo de la plataforma MiLab

Servicio	URL	VCPU'S (Cores)	Memoria (GB)	Disco (GB)	Sistema Operativo
chatLab	mattermost.redclara.net	2	4	50	Debian 10
compLab 1	jupyter.redclara.net	8	12	150	Debian 10
compLab 2	jupyterhd.redclara.net	2	4	50	Ubuntu 20.04
Despliegues	milabproy.redclara.net	2	2	50	Debian 10
dataLab	ckan.redclara.net	2	6	120	CentOS 7
G-Lab	gitmilab.redclara.net	4	8	50	Debian 10
IdP	idplaconga.redclara.net	1	2	50	Debian 10
LCMS	class.redclara.net	1	2	50	Debian 10
Web MiLab	milab.redclara.net	2	2	50	Debian 10
	Total	**24**	**42**	**620**	

Apéndice C. Ciencia Abierta

Existen diversas definiciones de la Ciencia Abierta, sin embargo, en este documento asumimos la definición creada por la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (en inglés, United Nations Educational, Scientific and Cultural Organization, UNESCO), en el marco de la *Recomendación de la UNESCO sobre la Ciencia Abierta*Educational et al. (2021). Un constructo inclusivo que combina diversos movimientos y prácticas con el fin de que los conocimientos científicos estén abiertamente disponibles y sean accesibles para todos, así como reutilizables por todos, se incrementen las colaboraciones científicas y el intercambio de información en beneficio de la ciencia y la sociedad, y se abran los procesos de creación, evaluación y comunicación de los conocimientos científicos a los agentes sociales más allá de la comunidad científica tradicional. Abarca todas las disciplinas científicas y todos los aspectos de las prácti-

cas académicas, incluidas las ciencias básicas y aplicadas, las ciencias naturales y sociales y las humanidades, y se basa en los siguientes pilares clave: acceso abierto al conocimiento científico, infraestructuras de la ciencia abierta, comunicación científica abierta, participación abierta de los agentes sociales y diálogo abierto con otros sistemas de conocimiento. UNESCO.

La Ciencia Abierta es una tendencia mundial, la evaluación e implementación de cada uno de sus elementos [Ver figura 25] aún tiene un gran camino por recorrer.

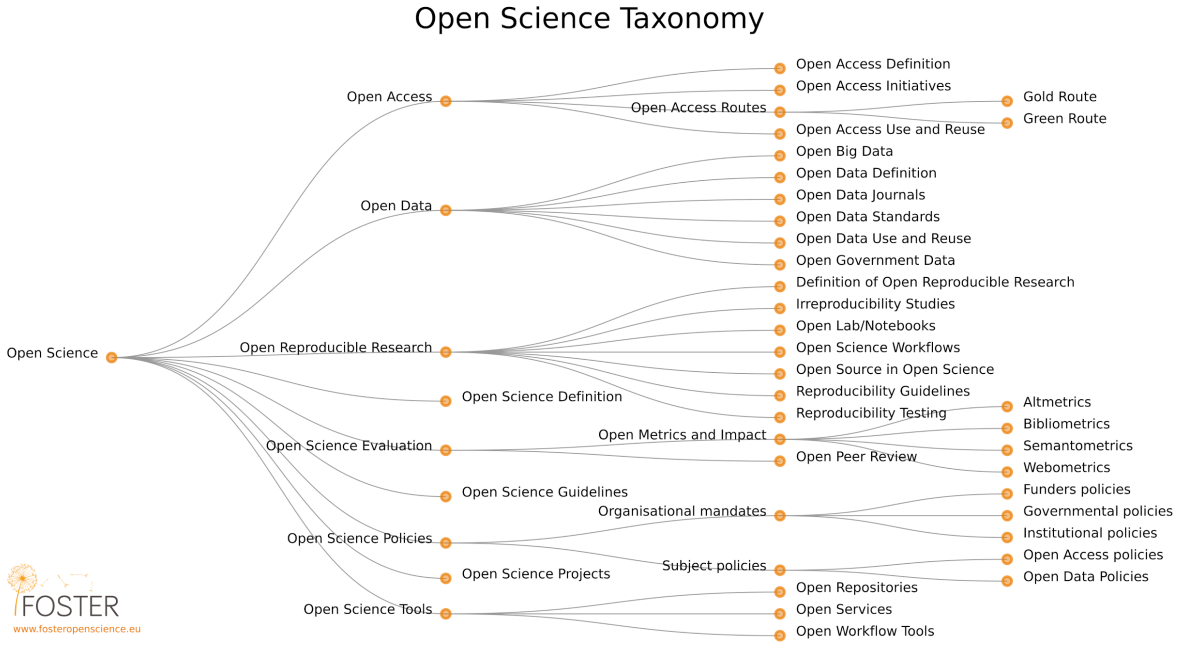

Figura 25. Taxonomía de la Ciencia Abierta. Adaptado de `https://doi.org/10.1145/2809563.2809571`